Gazi University Journal of Science
GU J Sci
24(4):829-839 (2011)

GUJS

Analysing Interactions of Risk Factors According to Risk Levels for Hemodialysis Patients in Turkey: A Data Mining Application

Yunus Y. ALTINTAŞ[1], Hadi GÖKÇEN[2], M. Mahir ÜLGÜ[3], Neslihan DEMİREL[2,♠]

[1] *The Undersecretariat of Treasury of Turkey, Directorate-General of Foreign Economic Relations, 06100, Ankara, Turkey*
[2] *Gazi University, Faculty of Engineering, Department of Industrial Engineering, 06570, Ankara, Turkey*
[3] *The Ministry of Health of Turkey Coordinatorship of Information Technologies, 06434 Ankara, Turkey*

Received: 10/02/2011 Accepted: 04/03/2011

ABSTRACT

The End Stage Renal Disease (ESRD) as a chronical health problem requires an expensive and a lifetime treatment called hemodialysis. It is important to obtain new information in order to reduce the cost of the treatment and to improve the quality of patient's life. Treatment period requires a lot of clinical tests related to the risk factors for monitoring patient's health and effectiveness of the treatment. These factors vary depending on demographic parameters such as age, gender, race, clinical parameters such as hematocrit level, albumine level, and also dialysis treatment prescription. In this paper, a data mining application including data preprocessing, data transformation, data mining algorithms and interpretation is used to find out patterns of risk factors as decision rules according to risk levels for dialysis patients in Turkey. A data set is formed by collecting 76 parameters of 170 patients on dialysis for 12 or more months at a dialysis center. CMS HCC (the Centers for Medicare and Medicaid Services -- Hierarchical Coexisting Conditions) ESRD model which includes relative coefficients of age, gender and comorbid diseases as scoring parameters is employed on data set in order to calculate risk scores for each patient and these scores are added to data set as a parameter called "Risk Score". ESTARD and WEKA softwares are used in order to achieve classification, clustering and decision tree algorithms. Decision rules as results of application are interpreted with domain expert for medical significance.

Keywords : *Hemodialysis, Risk level, Risk score, Data Mining.*

1. INTRODUCTION

In Turkey, approximately 45000 patients underwent dialysis at 13879 dialysis machines in 837 dialysis centers in 2008 [1]. The annual cost of dialysis patients exceeded 600 million USD in 2008. Due to the growing number of patients suffering from ESRD, it is estimated that more than 55000 patients will suffer from this disease in coming years in Turkey.

ESRD has become a frightening disease due to limited transplantation opportunities (600 transplantations in a year in Turkey, on average). It occurs when the kidneys are no longer able to function sufficiently and can not filter the toxins from the blood. It usually arises when chronic kidney disease has worsened to the point at which kidney function is less than 10% of normal. Patients who have reached this stage require dialysis or a kidney transplant [2]. Dialysis is a process that is performed routinely on people who suffer from acute or chronic kidney failure, or have ESRD. Dialysis involves removing

♠Corresponding author, e-mail: neslihanozgun@gazi.edu.tr

waste substances and fluid from the patient's blood, where this process is performed by the kidneys of a healthy person. There are two types of dialysis which are the peritoneal dialysis and hemodialysis [3].

Data mining techniques have the ability to observe the complex nature in data processes. The key to utilize the power of data mining techniques is to interpret and represent the derived knowledge in a meaningful way [4]. Although, data mining has been applied with success to different fields of human endeavor, its application to the analysis of medical data has until recently been limited [5]. Bellazzi and Zupan [5] reviewed the recent relevant work published in the area of predictive data mining in clinical medicine and gave guidelines to carry out data mining studies in this field. Kusiak et al. [6] monitored over 50 parameters to understand the collective role of these parameters in determining outcomes for an individual patient. Two different data mining approaches were employed to discover the knowledge between the measured parameters and patient survival. Bellazzi et al. [7] analyzed the data of more than 5800 dialysis sessions of 43 different patients and studied temporal data mining techniques for dialysis failure prediction. Knorr et al. [8] considered data mining in the medical settings of hemodialysis treatment and used a large dialysis treatment data set. They also provided a brief review of state-of-the-art methods for predicting patient risk and survival of dialysis patients.

In this paper, different decision tree algorithms are used to find patterns in a data set in order to analyze the interactions of risk factors according to risk levels for dialysis patients in Turkey. Raw data set is formed by collecting data of 170 patients on dialysis for 12 or more months, over 76 parameters. Factors which directly affect the death risk, are determined with the support of domain expert (nephrologist). After data transformation, a final data set is formed. Although they do not directly affect the death risk, some other parameters related to anticoagulation treatment, anemia treatment and arterial access type are also added to data set. Thus, it is aimed to gain knowledge about their interactions with directly risk related parameters.

2. MATERIAL AND METHODS

2.1. Data Collection

Raw data set is formed according to 76 different parameters, by the data collected from 170 patients on dialysis 12 or more months. Information of 170 patients' condition is summarized at Table 1. Collected parameters, except for patient ID, are separated to three main data groups as demographic parameters, dialysis prescription parameters and clinical parameters.

The demographic parameters are the patient's date of birth, start date of dialysis treatment, gender, age, blood type, the date of death, kidney transplant, transfer into dialysis center, transfer out from dialysis center, weight, height and blood pressure (systolic and diastolic) measured at treatment start, Body Mass Index (BMI), Tobacco and alcohol usage, diet type, the primary and secondary diagnoses.

The dialysis prescription parameters are composed of data recorded at dialysis sessions, as Pre-Session systol, Pre-Session diastol, Pre-Session weight, Post-Session systol, Post-Session diastol, Post-Session weight, Blood flow rate, Session frequency, Session duration, Anticoagulant dose, Inter-Session weight, Type of arterial access and in addition to these data Residive urine, Eritropoetin (EPO) pill type, EPO dose per week, Dialysis age, Lifetime at dialysis, Total session number. Average of last 15 sessions' measurements of dialysis prescription parameters is considered except for Dialysis age, Total session number and Session frequency [6].

Table 1. Information of 170 patients

Condition	Number	Ratio
At Hemodialysis	54	31,8
EX	68	40,0
Transfered	34	20,0
Transplanted	8	4,7
Started Periton Dialysis	2	1,2
Recovered	1	0,6
Unspecified	3	1,8
Total	170	100

The clinical parameters are collected from patients' test reports. For each parameter, average of 12-months period is considered in order to improve the data quality. The clinical parameters are formed for 1-month period measured Hb (Hemoglobin), Htc (Hematocrit), MCV (Mean Corpuscular Volume), MCH (Mean Corpuscular Hemoglobin), MCHC (Mean Corpuscular Hemoglobin Concentration), Leucosit, Platelet, Na (Sodium), K

(Potasium) (Input), K (Potasium) (Output), K (Potasium) Difference, Glycemi (Blood sugar level), Urea-BUN(Blood Urea Nitrogen) Input, Urea-BUN Output, Creatinine (Input), Creatinine (Output), URR (Urea Reduction Ratio), Kt/V (a calculated quantity to measure how well urea is removed in a dialysis session), Ca (Calcium), P (Phosphor), Ca x P, ALT (Alanine Transaminase), Total Protein, Albumine; for 3-months period measured CRP (C Reactive Protein), Uric Acid, PTH (ParaThyroid Hormone), Alkaline Phosphatase, Ferritine, Fe (Iron), Fe B.K. (Iron binding Capacity), TSI (Transferrin Saturation Index), Venous Bicarbonat; for 6-months period measured Total Cholesterol, HDL (High Density Lipoprotein), LDL (Low Density Lipoprotein) and Triglyceride.

2.2. Data Preprocessing and Transformation

The parameters collected from different databases are preprocessed by computing averages, ignoring records with missing data, filling in missing data and merged into an aggregate data set. Furthermore, data transformation is carried out by using domain knowledge, combining features and statistical methods. Before and after each session, the patient is weighted and their systolic and diastolic blood pressures are recorded. Total time for the dialysis session, session frequency, blood flow rate, blood volume, dialysis flow rate and anticoagulation dosage are adjusted depending upon the clinical and laboratory parameters for an individual patient. Required data are collected every month to measure the adequacy of dialysis assessed by the URR and Kt/V. The urea reduction rate (URR) is calculated by the difference in the blood urea concentration before and after dialysis divided by the pre-dialysis blood urea concentration [9]. The Kt/V is calculated as [(URR*0,023)-0,284] [10].

Current age is normally computed according to the year 2009, whereas for deceased, transferred or recovered patients, is computed by subtracting date of birth from the date of death/transfer or recovery. All values are expressed in years, except for the dialysis age which is expressed in months. For amputed patients (only one over 170 patients), weight is calculated by multiplying present weight with 1,15 [11].

Arterial access is a medical tool that connects patients to dialysis machine. As access type is important as a risk factor, eight different arterial access are added to data set for observation. Monthly measured clinical parameters of Ca and P are considered together as CaxP level for risk assessment of patients on dialysis. Since, TSI is selected as a significant parameter, parameters used to calculate TSI [which are Fe and Fe B.K.] are removed from the data set.

A new parameter called Weekly Session Duration is formed by multiplying Session frequency and Session Duration. Age at the start of dialysis is added to data set as a parameter. Other relevant parameters such as BUN Ratio, BMI, NRI (Nutritional Risk Index), ideal weight, post-session weight, EPO (Eritropoetin) resistance, EPO and anticoagulant dose are also calculated appropriately for data mining approach.

In order to improve data quality and classification accuracy, insignificant parameters such as Blood type, Session frequency, Session duration, Diet condition, Transplantation condition, Usage of tobacco/alcohol, Hemoglobin, MCV, MCH, MCHC, Leucosit, Platelet, Alkaline phosphatase, ALT (Alanine Transaminase) are removed from the data set in parallel to domain expert advise. Because of high diversity for nominalization, primer and seconder diagnoses of patients are also removed from the data set after used for risk scoring.

2.3. Risk Factors and Risk Scoring

In this study, CMS HCC ESRD model is used for risk assessment of ESRD patients [12]. CMS HCC Model was developed by Health Economics Research Inc. and it has been used legally in USA since January 2004 [13]. As a disease special version of model, CMS HCC ESRD model was developed in 2005 [14]. CMS HCC ESRD model uses age and gender as demographic data, primer and seconder diagnoses as clinical data and interactions of some diseases (if available) for risk assessment. Relative factors of patients are determined for age and gender conditions and are given in Table 2. Primer and seconder diagnoses are also used for risk scoring in model. It includes ICD-9 (International Classification of Diseases, 9th Revision) classification system for diseases and uses its clinic condition codes to identify diagnoses. Since relative factors for ESRD etiology are determined as 0 in ESRD Model, seconder diagnoses are the main parameters for risk scoring. In risk assessment of patients, relative factors of 2007 revision are used [15].

Table 2. CMS HCC ESRD Model Relative factors about age and gender

Age Groups	Relative Factors	
	Female	Male
0-34 Years	0,699	0,614
35-44 Years	0,699	0,650
45-54 Years	0,715	0,675
55-59 Years	0,746	0,699
60-64 Years	0,749	0,722
65-69 Years	0,813	0,776
70-74 Years	0,813	0,776
75-79 Years	0,831	0,790
80-84 Years	0,850	0,790
85 Years and Over	0,872	0,826

Seconder diagnoses of patients in the data set are classified according to related Disease Factor Groups based on valid ICD-9 codes. Relative factors and interactions of some disease groups, are given in Table 3. Sample Risk score calculations for patients is shown in Table 4.

Table 3. CMS HCC ESRD Model relative factors about primer and seconder diagnosis

HCC Code	Disease Group	Relative Factors
HCC1	HIV/AIDS	0,235
HCC10	Breast, Prostate, Colorectal and Other Cancers and Tumors	0,058
HCC18	Diabetes with Ophthalmologicor Unspecified Manifestation	0,080
HCC45	Disorders of Immunity	0,113
HCC80	Congestive Heart Failure	0,086
HCC108	Chronic Obstructive Pulmonary Disease	0,078
	Interactions	
INT1	DM_CHF(Diabetes+Chronic Heart Failure)	0,020
INT2	DM_CVD(Diabetes+Chronic Vascular Disease)	0,051
INT3	CHF_COPD(Chronic Heart Failure+Chronic Obstructive Pulmonar Disease)	0,000
INT4	COPD_CVD_CAD(Chronic Obstructive Pulmonar Disease+Cardiovascular Disease+Coronery Arter Disease)	0,000

Table 4. Sample Risk Score Calculation (Risk Score=Demographic Score + Primer Diagnose Score + Seconder Diagnose Score + Interactions Score)

ID	Age	Gender (F/M)	Primer Diagnose	Seconder Diagnoses				Interactions	Risk Score
				1	2	3	4		
1	72	M 0,776	Policistic Kidney Disease 0						0,776
2	46	M 0,675	Post Streptococcus GN 0	Hypertension 0,077					0,752
3	44	F 0,699	Post Streptococcus GN 0	Hypertension 0,077	CHF 0,086				0,862
4	62	M 0,722	Diabetic Nephropathy 0,08	Hypertension 0,077	Diabetes 0				0,879
5	56	M 0,699	Diabetic Nephropathy 0,08	Diabetes Type II 0	CHF 0,086	Blindness 0		DM-CHF 0,020	0,885

2.4. Data Mining Process

Methodology used to elicit knowledge about the interaction of risk factors according to risk levels for dialysis patients is given below:

STEP I: (Selection of key parameter as evaluation criteria)
As the first step of process, Risk Score is selected as a key parameter which is also an evaluation criterion of the data set. In order to get correct decision rules and interpret interactions of significant parameters according to risk levels, parameters which are significant and related to death risk are selected and insignificant parameters such as Blood Type, diet type, Transplantation state, Hb (Hemoglobin), MCV, MCH, MCHC, Leucosite, Platelet, Alkaline Phosphatase, ALT (Alanine Transaminase) are not taken into consideration as evaluation criteria. Tobacco and alcohol usage are removed from the data set due to inconvenience of collected data. Parameters like Session frequency, Session Duration, Ca, P, Fe (Iron) and Fe B.K. (Iron Binding Capacity) are also removed from the data set after they are used for data transformation. Information about parameters of final data set and profile of 170 patients on dialysis are given in Table 5.

STEP II: (Determination of class numbers for key parameter)
In this step, class numbers of key parameter Risk Score are determined. Clustering performance of data set is important in order to get correct decision rules from decision tree algorithms. It is required to determine the number of risk levels that Risk Score may have before applying clustering algorithm. Uniqueness ratio of parameters in the data set is key factor to make the decision of whether to add parameters to data mining process or not. ESTARD Data Miner is a data mining software which automatically computes uniqueness ratio of parameters and permits the selection of parameters that only have low uniqueness ratio for class formation. ESTARD Data Miner is used for determining different class numbers such as 2, 3, 4 and 5 for Risk Score. Information about determined Risk Score classes is given in Table 6.

Table 5. Information about parameters of final data set and profile of 170 dialysis patients

Parameter	Label	Unit	Min.	Max.	Average	Std. Deviation	Uniqueness Rate(%)
ID	ID	-	-	-	-	-	100
Gender	SEX	Nominal	**0:81, 1:89**				1,2
Age	AGE	year	21	86	59,41	15,99	33,5
Dialysis Start Age	SAGE	year	20	82	55,11	15,98	34,1
Weight	HEI	kg	145	186	163,74	8,62	19,4
Height	WEI	cm	40	99,7	65,38	12,01	60,0
Body Mass Index	VKI	kg/m2	13,9	39,3	24,06	4,39	100,0
Systol at Start	SS_AKB	mmHg	70	200	136,35	22,39	7,7
Diastol at Start	SD_AKB	mmHg	50	120	78,82	12,11	4,1
Nutrition Risk Index	NRI	-	30,4	80,1	51,05	8,53	100,0
# of Sessions	SES	number	86	2066	566,15	410,68	93,5
Total Treatment Time	SEAG	month	11	207	58,85	42,48	51,8
Time of Life in Dialysis	DHS	year	0,9	17,3	4,91	3,54	51,8
Session Duration/Week	HSS	hour	6	13,5	11,67	1,12	5,3
Arterial Access Type	DYT	-	**0:94,1:27,2:10,3:15, 4:3, 5:17, 6:1, 7:3**				5,3
Anticoagulant Dose	AKG	(unit/session)	0	10000	4406,35	1340,22	18,2
Pre-session Weight	Pre-W	kg	37,9	105,5	66,29	12,45	100,0
Pre-session Systol	Pre-S	mmHg	82	185	130,68	18,78	55,3
Pre-session Diastol	Pre-D	mmHg	52	103	76,80	8,49	31,8
Post-session Weight	Ps-W	kg	37,8	102,2	64,41	12,15	100,0
Post-session Systol	Ps_S	mmHg	77	172	121,08	19,90	53,5
Post-session Diastol	Ps-D	mmHg	50	93	72,84	8,56	35,3
Intersession Weight Diff.	ISW	kg	-1,4	4,2	1,88	0,89	95,3
Blood Flow Rate	KAH	Qb	180	300	244,96	22,76	12,9
Eritropoetin Drug Type	EIT	-	**0:102, 1:39, 2:20, 3:9**				2,4
EPO Dose	EPOD	(unit/month)	0	56000	20936,21	11881,59	79,4
Efficiency of EPO Dose	EPOU	(unit/month)/kg	0	1057,7	336,19	197,83	95,3
EPO Resistance	EPOR	-	0	108,6	32,70	21,33	95,3
Residive Urine	RZI	ml	0	4227	332,71	490,37	57,1
Hematocrite	Htc	%	25,92	44,59	32,44	3,50	98,8
C Reactive Protein	CRP	mg/l	0,08	64	4,72	9,22	68,8
Sodium	Na	mEg/dl	127,08	147,33	136,82	3,41	62,9
Potassium (Input)	K_I	mEg/dl	3,8	6,55	5,20	0,49	81,2
Potassium (Output)	K_O	mEg/dl	2,84	4,73	3,92	0,36	82,4
Potassium Difference	K_F	mEg/dl	0,45	2,32	1,27	0,41	90,0
Glycemi	GLI	mg/dl	73,83	278	121,72	49,58	93,0
Urea-BUN (Input)	BUN_I	mg/dl	65,17	213,67	140,90	28,49	95,3
Urea-BUN (Output)	BUN_O	mg/dl	18,42	91,67	51,57	11,34	90,0
BUN Ratio	BUN_R	-	1,94	4,15	2,77	0,39	99,4
Cretainin (Input)	KRE_I	mg/dl	2,36	15,96	8,28	2,25	94,1
Creatinin (Output)	KRE_O	mg/dl	1,27	6,73	3,84	1,04	92,4
Uric Acid	URA	mg/dl	3,93	9,4	6,33	0,91	78,2
Urea Reduction Ratio	URR	-	49,22	76,04	63,33	4,70	100,0
Urea Clearance Ratio	KtV	-	0,85	1,46	1,17	0,11	100,0
Calcium x Phosphor	CaxP	-	24,72	69,17	44,65	8,62	100,0
Parathyroid Hormone	PTH	pg/ml	12	1433,4	234,13	240,61	97,1
Total Chollesterol	Tkoll	mg/dl	104,5	310	177,72	36,01	73,5
High Density Lipoprotein	HDL	mg/dl	24	123,3	40,42	9,61	35,3
Low Density Lipoprotein	LDL	mg/dl	8	180,5	81,68	24,63	68,8
Triglyceride	TGL	mg/dl	53,5	982	227,71	114,02	92,4
Total Protein	TPRO	g/dl	5,41	7,88	6,80	0,40	77,7
Albumine	ALB	g/dl	2,84	8,54	3,94	0,50	76,5
Ferritine	FERT	mg/dl	40,2	2014,8	506,42	318,64	97,6
Transferrin Saturation Index	TSI	-	10,23	58,21	26,93	9,27	100,0
Risk Score	RSCORE	-	**0:104, 1:66**				46,5

Table 6. Information about Risk Score classes

# of Risk Score Class	Nominal Class Label	Class Interval	# of patients in Class	Ratio
2	0	0.614 - 0.870	104	61.2
	1	0.870 - 1.126	67	39.4
3	0	0.614 - 0.786	61	35.8
	1	0.786 - 0.955	85	50.0
	2	0.955 - 1.126	24	14.2
4	0	0.614 - 0.742	36	21.1
	1	0.742-0.870	68	40.0
	2	0.870 - 0.998	57	33.5
	3	0.998 - 1.126	10	5.4
5	0	0.614 - 0.716	31	19.0
	1	0.716 - 0.810	42	25.0
	2	0.810 - 0.920	59	34.0
	3	0.920 - 1.020	33	19.1
	4	1.020 - 1.126	5	2.9

STEP III: (Nominalisation of key parameter and formation of alternative data sets according to class numbers)
This step includes formation of alternative data sets by adding nominalised Risk Score column for class numbers 2, 3, 4, and 5. Selected clustering algorithm is applied to these data sets.

STEP IV: (Determination of the best class number for Risk Score and formation of the final data set)
K-means algorithm is used in order to compare clustering performances of data sets that have different class numbers for Risk Score. Clustering performances of data sets with 2, 3, 4, 5 classed Risk Score can be seen in Table 7. It can be concluded from the Table 7 that incorrectly clustered instance ratio is fewest for data set with 2 classed Risk Score, so its clustering performance is better than other data sets with 3,4 or 5 clustered Risk Scores. Therefore, decision tree algorithms are applied on final data set which is formed by adding 2 classed Risk Score to data set. For better interpretation of decision rules, Risk Score Class 0 is called as Risk Level 0 which is considered as low risk and Risk Score Class 1 as Risk Level 1.

Table 7. Clustering performances of data sets with 2, 3, 4 and 5 classed Risk Score

# of Class for Risk Score	k-means Algorithm			Average Ratio for Class
	# of Cluster	Incorrectly Clustered Instance		
		No	Ratio	
2	2	57	33,5	46,15
	3	64	37,6	
	4	94	55,3	
	5	99	58,2	
3	2	92	54,1	60,1
	3	95	55,8	
	4	109	64,1	
	5	113	66,4	
4	2	89	52,3	56,875
	3	89	52,3	
	4	102	60	
	5	107	62,9	
5	2	115	67,6	66,55
	3	111	65,2	
	4	111	65,2	
	5	116	68,2	

STEP V: (Determination of best parameters for estimations over the key parameter on the final data set)
Before the application of decision tree algorithms on the best clustering performance data set, it is required to determine the best parameters for estimations over the key parameter in order to check the decision rules.

CfsSubsetEval and Best first algorithms are applied on the data set with 2 classed Risk Score. As a result, 10 parameters (Age (AGE), Dialysis Start age (SAGE), gender (SEX), Height (HEI), Post-Session Diastol (Ps-D), C Reactive Protein (CRP), Potassium (Input) (K_I), Glycemi (GLI), Urea_BUN (Input) (BUN_I), Parathyroid hormone (PTH)) are determined as the best parameters for estimations over the key parameter (Risk score).

STEP VI: (Application of decision tree algorithms on final data set and evaluation of their performances)
Three different algorithms, J4.8, RandomTree and PART are selected as decision tree algorithms in order to apply over data set with 2 classed Risk Score. Performances of these algorithms are shown in Table 8.

Table 8. Performances of decision tree algorithms on data set with 2 classed Risk Score

# of Class for Risk Score	Decision Tree Algorithms					
	J4.8		RandomTree		PART	
	Incorrectly Classified Instances					
	No	Ratio	No	Ratio	No	Ratio
2	32	18,8	52	30,5	37	21,7

Results of J4.8 and PART algorithms are taken into consideration for the next step since their classification performances are better than Random Tree algorithm as seen in Table 8.

3. RESULTS AND DISCUSSION

STEP VII: (Results and Interpretation of Decision Rules)
The produced decision tree by J4.8 algorithm is given in Figure 1. Converted decision rules from the decision tree of J4.8 algorithm are given below:
1-If AGE<= 58 and GLI<= 155 Then RS0
2-If AGE<= 58 and GLI>155 and Htc<= 31.16 Then RS1
3-If AGE<= 58 and GLI>155 and Htc>31.16 Then RS0
4-If AGE>58 and SEX=1 and GLI<= 136.92 and WEI<=68.5 and DYT= 0 and HDL<= 37.5 Then7 RS1
5-If AGE> 58 and SEX=1 and GLI<=136.92 and WEI<=68.5 and DYT = 0 and
HDL>37.5 Then RS0
6-If AGE > 58 and SEX=1 and GLI<=136.92 and WEI<=68.5 and DYT=2 Then RS0
7-If AGE > 58 and SEX=1 and GLI<=136.92 and WEI<=68.5 and DYT=1 and ID<= 59 Then RS0
8-If AGE>58 and SEX=1 and GL<=136.92 and WEI<=68.5 and DYT=1 and ID>59
Then RS1
9-If AGE>58 and SEX=1 and GLI<=136.92 and WEI<=68.5 and DYT=5 Then RS0
10-If AGE>58 and SEX=1 and GLI<=136.92 and WEI<=68.5 and DYT=3 Then RS1
11-If AGE>58 and SEX=1 and GLI<=136.92 and WEI>68.5 Then RS0
12-If AGE>58 and SEX=1 and GLI>136.92 and K_F<=1.66 Then RS1
13-If AGE>58 and SEX=1 and GLI>136.92 and K_F>1.66 Then RS0
14-If AGE>58 and SEX=0 and SEAG<=91 Then RS1
15-If AGE>58 and SEX=0 and SEAG>91 Then RS0

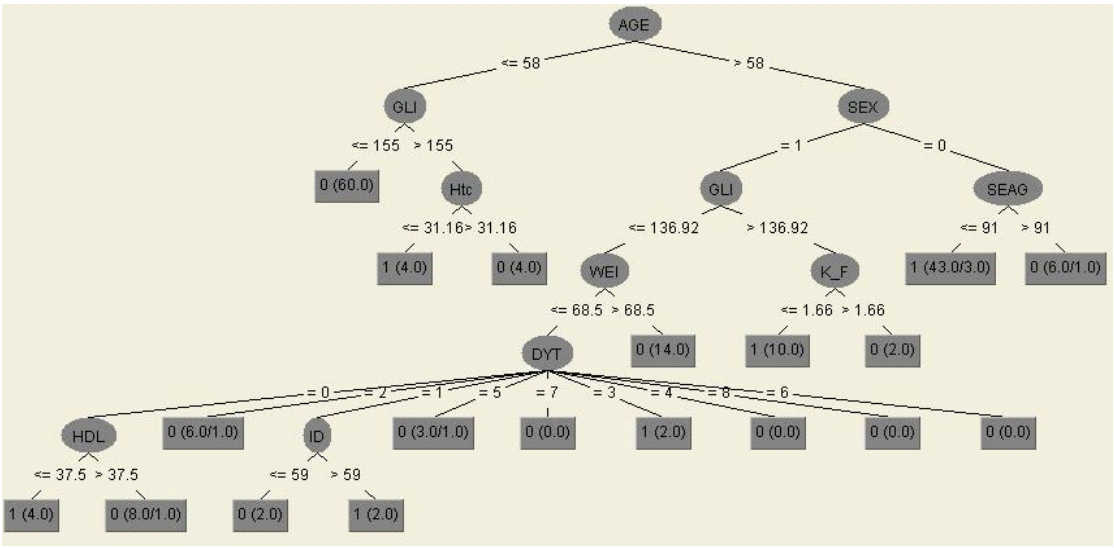

Figure 1. Decision tree of J4.8 Algorithm

Produced decision rules by PART Algorithm are given below:
1-If AGE <= 58 and GLI <= 155 Then RS0 (60.0)
2-If SEX = 0 and SEAG <= 91 and SEAG > 23 Then RS1 (38.0/1.0)
3-If K_I <= 4.83 and CRP > 0.58 Then RS1 (12.0)
4-If K_I <= 5.98 and Pre-S <= 142 and AKG <= 4423 and SES > 112 Then RS0 (15.0)
5-If DYT = 0 and TSI > 25.91 Then RS0 (11.0)
6-If EIT = 0 and SAGESEX <= 66 and TSI <= 19.44 Then RS0 (7.0)
7-If URA > 6.45 Then RS1 (9.0)
8-If Pre-S <= 142 and TKoll > 158 Then RS1 (8.0)

Decision Rules are interpreted as below:
* Patients of age 58 or more with Glycemi level above 155 mg/dl are in high risk group. Fasting blood sugar level (Glycemi) should be 70-110 mg/dl. If the value is below 70 mg/dl it is called Hypoglycemi, and over 110 mg/dl it is called Hyperglycemi. For the above rule, Glycemi level over 155 mg/dl and age over 58 means really high risk. Studies show that patients (within the age group of 55-64) with diabetes have shorter survival rate [6]. As a result, high age and high blood sugar level together increased the death risk.

* Second and third rules of J4.8 algorithm are "if age is 58 or below with Glycemi level over 155 mg/dl, Hematocrit level becomes important. Patients with %31,16 Hematocrit level and below have increased risk from level 0 to level 1". Hematocrit level should be between %33-36 in dialysis patients. According to rule, for patients aged 58 and below, high blood sugar level and low hematocrit level means increased death risk. As a result, together high blood sugar level and low Hematocrit level together increased the death risk.

*Fourth and fifth rules of J4.8 algorithm are "Male patients, aged above 58 years, having 68,5 kg or below body weight, using LAVF (Left Arteriovenous fistula) as arterial access type, having Glycemi level 136,92 mg/dl and below, HDL-Cholesterol (High Density Lipoprotein) level 37,5 mg/dl and below have increased risk from level 0 to level 1". Blood sugar level as 136,92 mg/dl and below is near normal considering 70-110 mg/dl interval. Glycemi level has no additive effect on death risk together with factors of HDL-Cholesterol and arterial access type – LAVF. Arterial access types are used according to patients' health condition and their effects on risk are unknown. Therefore, no comment about arterial access type parameter is stated, though it is situated in some correct rules. Low body weight decreases death risk for patients on dialysis but it is not convenient to comment on body weight parameter in rule. It could be more suitable to comment if body weight was in rule with BMI or NRI. Lower limit of HDL-Cholesterol level is 40 mg/dl for persons aged above 20 years. In above rules, HDL-Cholesterol level is 37,5 mg/dl and below. It means increased risk. Briefly, in these rules HDL–Cholesterol level has directly additive effect on risk and together with age and gender formed a meaningful pattern.

* According to second rule of PART algorithm "Female patients with treatment time between 23 months and 91 months has risk level 1". It is known that risk increases if treatment time gets longer. In this rule, there is a pattern about female patients and additive effect occurs on risk for patients who undergo dialysis more than 2 years till 7,5 years. As a result, gender and total treatment time together increased death risk.

* Third decision rule produced by PART algorithm is "Patients with Potassium (Input) level 4,83 mEq/L and below, C reactive protein level above 0,58 mg/l are classified in risk level 1". Blood Potassium (K) level of Patients who get dialysis treatment are measured two times a month regularly, before a hemodialysis session as K_I and at the end of the same session as K_O. Risk is increased below 3,5 mEq/L and above 6,5 mEq/L for Potassium Input and Output levels [16]. For C reactive protein level, it is determined that above 10 mg/l level means 3,5 times higher risk [17]. The interaction of K_I and CRP level increased the death risk.

* Fifth rule of PART algorithm is "Patients used LAVF as access type with TSI level above %25,91 have risk level 0". The most important reason for Eritropoetin resistance is iron deficiency. Therefore, TSI level should be at least %20 for hemodialysis patients [18]. As seen at rule, TSI level above %25,91 means lower death risk.

* Seventh rule produced by PART algorithm is "Patients with uric acid level above 6,45 mg/dl are in high risk group". Upper limit for uric acid level in blood is 6-7 mg/dl for women and 7-8 mg/dl for men [19]. Uric acid level above 6,45 mg/dl means increased death risk.

4. CONCLUSION

Data mining techniques have the ability to extract hidden knowledge in the data set and interpret the derived knowledge in a meaningful way. It is obvious that new tools are required to capture complex relationships between treatment, medications, and patient specific factors. There are some data mining studies applied in medicine. However, to the best of our knowledge, this study is the first study done on hemodialysis patients in Turkey. In this study, data mining approach is used to gather new risk patterns from hemodialysis patients' data.

Decision rules produced by J4.8 and PART algorithms showed that 5 of 10 best parameters for estimations about Risk score exist in rules, which are age (AGE), gender (SEX), C Reactive Protein (CRP), Potassium (Input) (K_I) and Glycemi (GLI). Interactions that increased the death risk are determined as age and glycemi, gender and total treatment time, glycemi and Hematocrit level, Potassium (Input) level and C Reactive Protein level and Age-gender and HDL-Cholesterol level. C Reactive Protein level and Uric acid level are determined as solely risk additive parameters according to decision rules.

Since reaching more valuable knowledge depends on analysis of huge amount of data, using a large data including all dialysis centers in Turkey will be an opportunity to invent medical discoveries of new treatment, new medicine and more useful guidelines, and enhance the quality of the patients' life.

ACKNOWLEDGEMENTS

Our special thanks to Dr. Eyup OZEREN, Assoc. Prof. Dr. Metin YILDIRIMKAYA and Ömer ASLANTAS for their support on data collection, Prof. Dr. Müjdat

YENICESU for his guidance on medical issues and Dr. Ali Rıza ODABAS, Dr. Hadim AKOGLU for their guidance on risk scoring.

REFERENCES

[1] Internet: DIADER Report, "The Role of Private Dialysis Centers on Hemodialysis Service in Turkey, Relation of Quality-Cost, Costs and Paybacks of Dialysis, Reducing State Costs", http://www.diader.org.tr/dosya/DiyalizRaporEkim2009.pdf, October (2009).

[2] Internet: http://www.nlm.nih.gov/medlineplus/ency/article/000500.htm M.D. Parul Patel, End-stage kidney disease December 8 (2009).

[3] Internet: End Stage Renal Disease, http://www.healthsystem.virginia.edu/UVAHealth/adult_urology/endstage.cfm September 18 (2007).

[4] Mries, M.F., "Modeling of hemodialysis patient hemoglobin: a data mining exploration" Master Thesis, ***The University of Iowa,*** (2007).

[5] Bellazzi, R., Zupan, B., "Predictive data mining in clinical medicine: Current issues and guidelines", ***International Journal of Medical Informatics,*** 77, 81-97 (2008).

[6] Kusiak, A., Dixon, B., Shah S., "Predicting Survival Time for Kidney Dialysis Patients: A Data Mining Approach", ***Computers in Biology and Medicine*** 35, 311–327 (2005).

[7] Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R., Temporal data mining for the quality assessment of hemodialysis services, ***Artificial Intelligence in Medicine,*** 34 25-39 (2005).

[8] Knorr, T., Schmidt-Thieme, L., Johner, C., Identifying Patients at Risk: Mining Dialysis Treatment Data in: Okada, A., Imaizumi, T., Bock, H.H., Goal, W., eds., ***Cooperation in Classification and Data Analysis,*** Springer Berlin Heidelberg, 131-140, (2009).

[9] Daugirdas, J.T., Evanstone, J.C., Physiological principles and urea kinetic modeling, in: Daugirdas, J.T., Ing, T.S., eds. ***Handbook of Dialysis***, Boston, Little Brown 15-45, (2000).

[10] Basile, C., Casino, F., Lopez, T., "Percent reduction in blood urea concentration during dialysis estimates Kt/V in a simple and accurate way", ***American Journal of Kidney Diseases,*** 15,40-45 (1990).

[11] Internet: Implementation of Changes in End Stage Renal Disease (ESRD) Payment for Calendar Year 2006, http://www.cms.gov/MLNMattersArticles/downloads/MM4196.pdf, (2006).

[12] Buntin, M.B., Garber, A.M., McClellan, M., Newhouse, J.P., "The Costs of Decedents in the Medicare Program: Implications for Payments to Medicare+Choice Plans", ***Health Services Research,*** 39(1), 111–130 (2004).

[13] Levy, J.M., Robst, J., Ingber, M.J., "Risk-Adjustment System for the Medicare Capitated ESRD Program", ***Health Care Financing Review*** 27(4), 53-69 (2006).

[14] Pope, G.C., Kautter, J., Ellis, R.P., Ash, A.S., Ayanian, J.Z., Iezzoni, L.I., Ingber, M.J., Levy, J.M., Robst, J., "Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model", ***Health Care Financing Review,*** 25(4), 119-141 (2004).

[15] Internet: Announcement of Calendar Year (CY) 2008 Medicare Advantage Capitation Rates and Payment Policies**,** http://www.cms.hhs.gov/MedicareAdvtgSpecRateStats/Downloads/Announcement2010.pdf April 2nd, (2007).

[16] Factor, K.F., "Potassium Management in Pediatric Peritoneal Dialysis Patients: Can a Diet With Increased Potassium Maintain a Normal Serum Potassium Without a Potassium Supplement?", ***Advances in Peritoneal Dialysis,*** 23, 167-169 (2007).

[17] Korevaar, J.C., van Manen, J.G., Dekker, F.W., de Waart, D.R., Boeschoten, E.W., Krediet, R.T., "Effect of an Increase in C-Reactive Protein Level during a Hemodialysis Session on Mortality", ***Journal of the American Society of Nephrology,*** 15(11), 2916-2922 (2004).

[18] Evrenkaya, T.R., Atasoyu, E.M., Ünver, S., Gültepe, M., Narin, Y., Tülbek, M.Y., The relationship between hemodialysis adequacy and co-morbid factors, ***Turkish Nephrology Dialysis and Transplantation Journal*** 2(1), 44 (2002).

[19] Obermayr, R.P., Temml, C., Gutjahr, G., Knechtelsdorfer, M., Oberbauer, R., Klauser-Braun, R., "Elevated Uric Acid Increases the Risk for Kidney Disease", ***Journal of the American Society of Nephrology,*** 19(12), 2407–2413 (2008).