



ELSEVIER

Computers in Biology and Medicine 35 (2005) 311–327

<http://www.intl.elsevierhealth.com/journals/cobm>

Computers in Biology
and Medicine

Predicting survival time for kidney dialysis patients: a data mining approach

Andrew Kusiak^{a,*}, Bradley Dixon^b, Shital Shah^a

^a*Intelligent Systems Laboratory, College of Engineering, 3131 Seamans Center, Iowa City, Iowa 52242 1527, USA*

^b*University of Iowa Hospital and Clinics, E300D GH, The University of Iowa, Iowa City, Iowa 52242 1081, USA*

Received 25 September 2003; accepted 17 February 2004

Abstract

The cost for providing care for patients on hemodialysis due to end stage kidney disease is high. Finding ways to improve patient outcomes and reduce the cost of dialysis is important. Dialysis care is intricate and multiple factors may influence patient survival. Over 50 parameters may be monitored on a regular basis in providing kidney dialysis treatments. Understanding the collective role of these parameters in determining outcomes for an individual patient and administering individualized treatments allowing specific interventions is a challenge. Individual patient survival may depend on a complex interrelationship between multiple demographic and clinical parameters, medications, medical interventions, and the dialysis treatment prescription. In this research, data preprocessing, data transformations, and a data mining approach are used to elicit knowledge about the interaction between many of these measured parameters and patient survival. Two different data mining algorithms were employed for extracting knowledge in the form of decision rules. These rules were used by a decision-making algorithm, which predicts survival of new unseen patients. Important parameters identified by data mining are interpreted for their medical significance. The concepts introduced in this research have been applied and tested using data collected at four dialysis sites. The computational results are reported in the paper.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: Hemodialysis; Survival; Data mining; Data preprocessing; Data transformations; Decision making; Medical relevance; Dialysis protocol

* Corresponding author. Tel.: +1-319-335-5934; fax: +1-319-335-5669.

E-mail addresses: andrew-kusiak@uiowa.edu (A. Kusiak), bradley-dixon@uiowa.edu (B. Dixon), scshah@engineering.uiowa.edu (S. Shah).

URL: <http://www.icaen.uiowa.edu/~ankusiak>

1. Introduction

Approximately 370,000 Americans undergo dialysis, at an annual cost of \$11.1 billion [1]. More than 260,000 Americans suffer from chronic renal failure and around 50,000 people die each year due to kidney failure [1,2]. Little attention is being paid to kidney health, despite the fact that the number of Americans with kidney failure is growing at a rate of 6% a year with the US leading the world in the number of new cases per million population [1,2]. In 1997, more than 79,000 Americans developed end-stage renal disease, bringing the total number of Americans treated for end-stage renal disease to more than 360,000 [2]. Chronic renal failure occurs when the kidneys are operating at less than 50% of normal capacity [3]. End-stage renal disease (ESRD) occurs when the kidneys are working at less than 10%–15% of normal capacity [3,4]. At this stage, either transplantation or repetitive kidney dialysis becomes necessary for survival.

Hemodialysis (HD) and peritoneal dialysis are the two modalities of dialysis treatment. HD is typically performed in a clinic setting and accounts for more than 80% of the dialysis population. Peritoneal dialysis accounts for the remaining 20% and is usually performed by the patient at his/her residence. During HD, the blood passes through an extra-corporal circuit where metabolites are eliminated, the acid–base equilibrium is re-established and excess salt and water is removed [5]. The process of diffusion exchanges solutes and metabolites across a semi-permeable membrane, separating the blood and dialysate [6]. Water is removed from the body using a negative pressure gradient in a process called ultra-filtration. After transit through the dialyzer, the clean, filtered blood is returned to the body. A device called a hemodialyzer regulates the entire procedure. Typically, HD is performed three times a week for 3–4 h each session, but dialysis time for these sessions varies from patient to patient [4,7]. Peritoneal dialysis works on the same principles of solute diffusion and fluid ultrafiltration as HD, but the blood is cleaned inside the body rather than through a machine. The kidneys filter nearly 189 liters of liquid from blood per day but only about 1% (1.89 l) of the original filtrate ultimately appears in the final urine as waste products and extra water. The waste products are not reabsorbed and are concentrated in the final urine. These waste products such as urea and creatinine are derived from the normal breakdown of foods and tissues. The kidneys also maintain the stability of the extracellular fluid (ECF) volume and electrolyte homeostasis by adjusting excretion of water and electrolytes to balance changes in intake [5]. In addition to these excretory functions, the kidney is an endocrine organ that produces hormones such as erythropoietin needed for red blood cell production and metabolizes vitamin D into an active form needed for proper bone growth and turnover. The kidneys are also the primary route for elimination of many foreign substances such as drugs, food additives, pesticides, and other components from the body [5]. With kidney failure, waste products build up in the body, fluid and water homeostasis is impaired and the endocrine functions of the kidney are deranged. This impairs the function of multiple organ systems producing a toxic condition known as uremia that if not corrected will lead to death [5].

Although dialysis is life saving for a person with terminal kidney failure, survival is still markedly reduced compared to an age-matched healthy population. The median life span of a dialysis patient in the US is slightly more than 3 years [8]. Kidney failure due to health conditions such as diabetes accounts for much of this excess mortality. The description of known significant parameters, their interaction and some of the indexes are included in the Medical Significance Section 6. However, the observation that recipients of kidney transplant do better than similar patients who receive dialysis while on the waiting list for a kidney transplant suggests that the process of dialysis could be

improved. Targeted interventions for high-risk patients and improvements in the dialysis prescription are possible by understanding factors that are predictive of survival of a given patient.

Survival analysis using standard statistical tools, such as logistic regression, Cox's model, factorial designs, and so on can be considered as "population-based" models. Predictions are derived on probability or distance from the population estimates. Data mining offers tools for decision-making for an individual patient rather than a population of patients. It provides tools for identifying valid, novel, potentially useful, and ultimately understandable patterns from data and constructs high confidence predictions for individuals [9,10]. Discovering hidden patterns in data may represent valuable knowledge that can lead to medical discoveries, e.g., certain ranges of parameter values that lead to a longer survival time. In this paper, a data mining approach was used to identify relevant factors contained in a data set of routinely collected clinical and hemodialysis parameters that are predictive of an individual patient surviving beyond the median survival time.

2. Data collection and preprocessing

Data collection was performed at four satellite locations of The University of Iowa Hospitals and Clinics (UIHC) in a routine process and was provided for this research project. Data collection was based on known as well as unknown indicators (parameters) of effectiveness of dialysis treatment. Dialysis patients received the treatment three times a week. Before each session, the patient is weighed and their blood pressure registered while they are sitting (supine pressure), and when possible, while they are standing. The systolic and diastolic blood pressure measurements are recorded. The weight and blood pressure measures are repeated at the end of the session. The levels of sodium, bicarbonate, potassium, calcium, and glucose in the dialysis solution are recorded. They are adjusted depending upon the clinical and laboratory parameters in an individual patient. Total time for the dialysis session, blood flow rate, blood volume and dialysis flow rate, were another set of collected measures. The type of dialysis machine and the overall average arterial and venial pressures are recorded as well. There is a set of measurements collected by the machine every 20 minutes as well as every time a nurse requests it. This set includes blood pressure, pulse, blood flow rate, arterial and venial pressures, trans-membrane pressure, and the rate of ultrafiltration. The machine can save 15 sets of these measures during each dialysis session. Another dialysis data set consisting of access flow rate is obtained approximately once a month. The information included in this set is the type of access (either arteriovenous fistula, which is natural, or an arteriovenous graft composed of Gore-tex, which is artificial), the location of access (right or left arm), and the average flow rate. This data is important because if the average flow rate is low, it means that the patient's access is beginning to narrow and a procedure will be needed to correct the stenosis. Patients who receive a catheter for dialysis rather than an arteriovenous fistula or graft do not have monthly measurements of access flow.

Once per month data is collected to measure the adequacy of dialysis assessed by the urea reduction ratio and Kt/V (a calculated quantity to measure how well urea is removed in a dialysis session). This data set included the blood urea nitrogen level before and after the dialysis session, the total time of the session, the patient's weight before and after the session, the target weight and body surface area. The urea reduction rate (URR) was calculated from the difference in the blood urea concentration before and after dialysis divided by the pre-dialysis blood urea concentration. The Kt/V

Table 1
Computing derived parameters

Derived blood pressures (orthostatic)		
	Pre	Post
Systolic (S)	Pre PS–Pre TS	Post PS–Post TS
Diastolic (D)	Pre PD–Pre TD	Post PD–Post TD
Derived blood pressures (time dependent)		
	Supine (P)	Standing (T)
Systolic (S)	Post PS–Pre PS	Post TS–Pre TS
Diastolic (D)	Post PD–Pre PD	Post TD–Pre TD
Derived pulse pressure		
	Supine (P)	Standing (T)
Pre	Pre PS–Pre PD	Pre TS–Pre TD
Post	Post PS–Post PD	Post TS–Post TD

was calculated using the URR and adjusted for fluid shifts that occur during dialysis. Finally, the normalized protein nitrogen appearance (nPNA) measuring the adequacy of the patient's nutritional intake was calculated. The data set contained missing values.

The demographic and outcomes data set contained the patient's date of birth, gender, and race; the date(s) of death, kidney transplant, transfer into dialysis center, and transfer out of dialysis center, and the diagnosis codes for the primary and secondary diagnoses. The data from the dialysis machines was organized by individual visits. To reduce data noise, averages were computed over the 15 readings taken by the machine during the dialysis session. Averages were computed for the time of each reading and all nine parameters recorded by the machine.

Data from the dialysis machines was organized to portray summaries of each patient's information. A data set with 188 patients was created by combining all four outreach centers. The number of dialysis sessions ranged from one to 707 visits. In initial analysis the data for all patients was used, and then a sub-group of long-term dialysis patients was considered. The pulse pressure (the difference between systolic and diastolic pressures) and the change in blood pressure during the dialysis session were of interest to medical experts. In order to address these parameters, the derived parameters shown in Table 1 were computed.

For those patients who are still alive, data such as current age, age at the start of dialysis, and their total time on dialysis was obtained. An "A" to represent "alive" was entered in the survival time column. For deceased patients, the analysis was performed using age at the start of dialysis, age at death, and total survival time after starting dialysis treatments. A "D" to represent "deceased" was entered in their current age column. All values were expressed in years, except the total survival time, which was expressed in months. Survival time is expressed as an integer as needed by the data-mining algorithm used in the paper. The Kt/V data files for all centers were combined, and then averages for each patient were calculated for each parameter in the data set. The number of readings for each patient ranged from 0 to 47. The Kt/V data set was very consistent with only a few missing data points.

The ideas proposed in this paper have been developed and tested using distributed databases containing information on patients undergoing dialysis. The databases were residing at four different

locations. They contained over 500 parameters for hundreds of patients, however the data was sparse. The data for some 50 different parameters (categorical, normative, and quantitative) was relatively complete.

3. Data transformation

Most data mining algorithms establish associations among individual parameter values. The approach used in this paper captures relationships among parameter functions. The concept of a parameter function is illustrated in Example 1.

Example 1. Consider the “as-is” data set in Fig. 1(a) with four parameters F1–F4, decision D, and five objects. Classification quality of a parameter set can be expressed as the percentage of all objects in a data set that can be unambiguously associated with the decision value based on this parameter set [11]. The classification quality of the parameters in Fig. 1(a) is as follows: $CQ(F1)=0.2$, $CQ(F2)=0.4$, $CQ(F3)=0$, $CQ(F4)=0.6$. The data set in Fig. 1(a) was transformed in the data set of Fig. 1(b), where the two parameters F2, F4 were replaced with the parameter function F2_F4.

The classification quality of the parameter sequence F2_F4 has the value $CQ(F2_F4)=0.6$, which is higher than that of individual parameters F2 and F4. The one-out-of $n(n = 5)$ cross-validation scheme [12] has been applied to the rules generated from the data sets in Figs. 1(a) and (b). The results of cross-validation are presented in Fig. 1(c) and (d). The average classification accuracy has increased from 20% for the rules extracted from the data set in Fig. 1(a) to 60% for the transformed data in Fig. 1(b).

Example 1 illustrates one of many possible parameter functions. Parameter sequences with a larger number of parameters have been successfully tested on a large-scale data set involving equipment maintenance [13]. The introduction of sequences in the maintenance data set has improved classification accuracy by 20%. The research team’s computational experience and the results published in the literature [14–16] indicate that the classification accuracy of the decision rules involving relationships between parameter functions may exceed those of the traditional decision rules. In addition,

No.	F1	F2	F3	F4	D
1	0	1	0	2	0
2	1	1	0	2	2
3	0	0	0	0	0
4	0	1	1	1	1
5	0	0	1	3	0

(a)

No.	F1	F2_F4	F3	D
1	0	1_2	0	0
2	1	1_2	0	2
3	0	0_0	0	0
4	0	1_1	1	1
5	0	0_3	1	0

(b)

	Correct	Incorrect	None
Average	20%	60%	20%

(c)

	Correct	Incorrect	None
Average	60%	20%	20%

(d)

Fig. 1. Data transformation example: (a) Data set with four parameters; (b) Transformed data set of Fig. 1(a), (c); Classification accuracy for the data set in Fig. 1(a), (d); Classification accuracy for the transformed data set of Fig. 1(b).

forming parameter functions according to user preferences may enhance the transparency of decision-making.

The number of transformations were performed on the following parameters used in this research: Pre_Supine_S, Pre_Supine_D, Pre_Supine_MAP, Pst_Supine_S, Pst_Supine_D, Pst_Supine_MAP, Diff_Pst_PreSupine_S, Diff_Pst_PreSupine_D, TBV, TD, BFR, DFR, NA, K, Diff_Pst_PreStand_S, Diff_Pst_PreStand_D, Diff_S_P_Supine_Pre, Diff_S_P_Supine_Post, Diagnosis, and Age_Begin. In addition to creating derived parameters, the data was discretized into 10 intervals and the pairs of parameters (parameter functions) have been constructed for the following parameters: Pre_Supine_S and Pre_Supine_D, Pre_Supine_MAP and Pst_Supine_MAP, Pre_Stand_MAP and Pst_Stand_MAP, Pulse_Supine_Pre and Pulse_Supine_Post, Diff_Pst_PreSupine_S and Diff_Pst_PreSupine_D, Pre_Supine_MAP, and TBV.

4. Decision making

Three decision categories were created for data mining purposes. The category “Above-median” consisted of patients who have either survived more than 3 years on dialysis and are still alive or deceased. The second category “Below-median” consisted of patients who survived less than 3 years on dialysis. The third category “Undetermined” consisted of patients who are alive and have not yet completed 3 years on dialysis. The group of patients in the “Undetermined” category could become either of the two categories and was excluded from data mining.

4.1. Data mining

In order to improve classification accuracy, insignificant parameters and patient data were removed from the data set. For example, all patients with less than 15 visits were removed from the data set.

The data for patients who received transplants required special consideration. Transplant’s success depends on factors such as age, prior time on dialysis, gender, primary cause of the disease, and so on [1]. In traditional data analysis, data for these patients would be censored following the transplants. However, in this dialysis data set, there is only one record per patient. Censorship would imply that data for patients who received transplants would have to be omitted from mining, thus making the data set too small for analysis. Therefore, the dialysis time for each transplant patient was calculated using the dialysis beginning date and the transplant date. The patients with dialysis time (before the transplant) greater than 3 years were classified as above-median while others were classified as below-median. There is no way to determine whether the below-median patients (with the transplant) would have survived on dialysis for more than 3 years. Due to the small data set, it was assumed that without transplant these patients would have died before the median survival time. Also, it was assumed as they were placed on the transplant list, their condition was deteriorating and the chances of them surviving above-median were small. The data set also contained data for four patients who had returned to dialysis treatment following a transplant. Since this situation is rather unusual their data records were removed from the data set.

All initial data mining was conducted using a rough-set (RS) algorithm. In the RS theory lower and upper approximations of the concept are computed [11]. As a result, there are two types of

Table 2
Test data sets

Trial set	Parameter description
T1	All continuous values
T2	All discrete values
T3	Nearly all values included
T4	Excludes all blood pressures except the continuous value differences
T5	Excludes all blood pressures except the discrete value differences
T6	All continuous blood pressure values except differences
T7	Continuous chemical values only
T8	Discrete chemical values only

decision rules, *certain* or *approximate*. *Certain* rules are induced from the lower approximations. The *approximate* rules are induced from the upper approximation, where upper approximation refers to the set of observations that can be possibly classified into this concept [11]. Classification accuracy is determined by dividing the number of objects in all lower approximations by either the number of objects in all upper approximates or all objects in the data set [11]. A decision-tree (DT) algorithm was an alternative data-mining tool to analyze the dialysis data. The DT algorithm creates a decision trees or sets of decision rules [17] based on the concept of information gain. An elaborate description of the RS and DT algorithms can be found in [11,17].

Sixteen different classifiers were generated by two data mining algorithms from eight sub-data sets created from the master data set. Thereafter, a decision-making algorithm was developed for predicting outcomes. Toward the end of this research, some patients in the “undetermined” category have become the above- or below-median category, and they were used for verification and validation of the decision-making algorithm.

4.2. Decision-making algorithm

Eight different test data sets were created (Table 2). The rough-set theory and DT algorithms were applied to eight data sets, producing 16 classifiers. These classifiers were used to make predictions. Each classifier can be considered as a digital expert. If all digital experts generate the same outcome, then this would be a confident prediction. The predictions generated by the 16 classifiers are often not identical, and therefore a decision-making algorithm is needed (see Fig. 2).

A simple voting scheme was used with each classifier having one vote. Thus the decision outcome with maximum number of votes is the predicted outcome. A conservative approach was applied while handling cases where there was a tie. The outcome “Below-median” was assigned, meaning that the patient would not survive above 3 years. A similar approach was implemented for each classifier to arrive at the final decision for that classifier (Table 3). Thus each rule within a particular classifier was provided with one vote. The outcome for each rule was either below-median, above-median or not applicable. In Table 3, Case 1 matches rules for both above- and below-median with equal votes, thus a conservative decision of below-median is taken. Case 9 on the other hand matches only the two below-median rules. Thus the below-median decision is assigned.

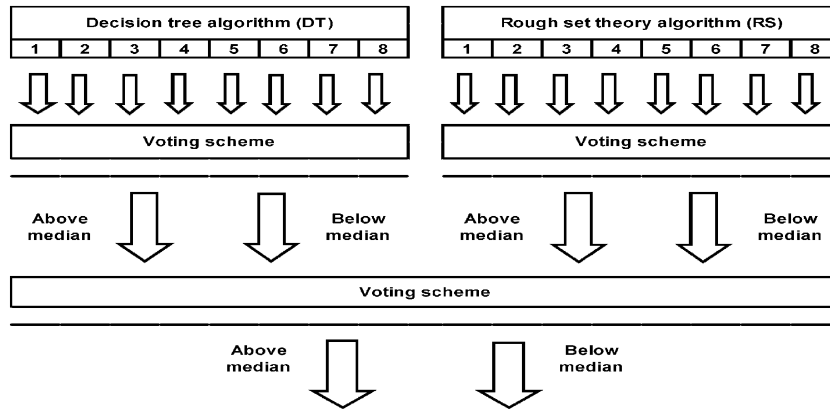


Fig. 2. Decision-making scheme.

Table 3
Sample of simple and weighted voting scheme

Rule	R1	R2	R3	R4	R5	R6	R7	R8	Simple		AD	ED	Weighted		AD	ED	Decision type	
	Case/CA	22	37	20	30	30	26	24	20	BM	AM			BM	AM			
1	BM	N	N	N	N	N	N	N	AM	1	1	BM	?	22	20	BM	?	Approx.
2	N	N	N	N	N	AM	N	AM	0	2	AM	AM	0	46	AM	AM	Exact	
3	N	BM	N	N	N	N	N	N	1	0	BM	BM	37	0	BM	BM	Exact	
4	N	N	N	N	N	N	N	N	0	0	?	?	0	0	?	?	No Rule	
5	BM	N	N	N	N	N	N	N	1	0	BM	BM	22	0	BM	BM	Exact	
6	N	N	N	N	N	N	N	N	0	0	?	?	0	0	?	?	No Rule	
7	N	N	BM	N	N	AM	N	AM	1	2	AM	?	20	46	AM	?	Approx.	
8	N	BM	N	N	N	N	N	N	1	0	BM	BM	37	0	BM	BM	Exact	
9	BM	BM	N	N	N	N	N	N	2	0	BM	BM	59	0	BM	BM	Exact	

CA: Classification Accuracy
AD: Approximate Decision
BM: Below_med
ED: Exact Decision
AM: Above_med
N: Not applicable

Parameters considered for generating classifiers may not be associated with the decision and they may be medically insignificant. Also, the quality of rules defining the classifiers may vary. The later two were taken into account by a weighted voting scheme. In this scheme the rule strength of a rule was used as weight and the weights were added if more then one rule matched the same outcome. Thus the decision outcome with maximum number of weighted votes is the predicted value. The same conservative approach as above was applied in case of a tie. Similar technique was implement within each classifier to arrive at the final decision for that classifier. Each rule was assigned a weighted vote (Table 3). The outcome for each rule was either below-median, above-median or not applicable. For example, Case 1 in Table 3 matches rules for both the above- and below-median survival, however, with different rule strengths thus resulting in the below-median decision.

Table 4
Quality measures

Decisions	one-out-of n			10-fold		
	Classification accuracy			Classification accuracy		
	Correct	Incorrect	None	Correct	Incorrect	None
<i>Cross-validation results: "As-is" data set</i>						
Below_med	45.24	54.76	0	23.93	72.74	3.33
Above_med	93.06	6.94	0	93.89	6.11	0
Average	75.44	24.56	0	67.65	31.44	0.91
<i>Cross-validation results: transformed data set</i>						
Below_med	64.29	35.71	0	52.83	47.17	0
Above_med	95.83	4.17	0	95.75	4.25	0
Average	84.21	15.79	0	75.97	24.03	0

5. Computational results

5.1. Algorithms

The RS theory algorithm generated rules of higher confidence, and the overall misclassification was lower. A sample rule generated by the RS algorithm is shown next.

```
IF (Diff_Pst_PreSupine_S < -3.81016) AND (Diff_Pst_PreStand_S
    > = -21.7308) AND (Ca < 3.40226) THEN (Survival_Length
    = Above_med)
```

The DT algorithm generated rules containing fewer parameters than the rules produced by the RS algorithm. The version of the DT algorithm used in this research had a property of assigning a default decision if none matched, and the overall misclassification rate was a bit higher. A sample DT rule is shown next.

```
IF (d.NA = 9) AND (Ca < = 3.346535) AND (d.Art.S = 6) THEN
    (Survival_Length = Above_med)
```

5.2. Data mining results and analysis

The one-out-of n (1 out of 114) and k ($=10$) fold (10 times 10% out of 114) cross-validation results for "as-is" data set with 114 patient data set are shown in Table 4. The one-out-of n cross-validation produced higher classification accuracy with 28 misclassifications (5 for above-median and 23 for below-median). The above median outcome was predicted with the accuracy of 93% for both one-out-of n and 10-fold cross-validation schemes.

Table 5
Comparison of results produced by the DT and RS algorithms

Rule set summary			Parameter occurrences					
Rules	DT	RS	Parameters	DT	RS	Parameters	DT	RS
<i>Trail set 4</i>								
Number of rules	16	15	Off.Target		2	TBV		1
Type of rules	Approximate	Absolute	Diff_Pst_PreSupine_S		1	TD		5
			Diff_Pst_PreSupine_D	5	3	BFR	1	3
Cases described			Diff_Pst_PreStand_S	1	2	DFR		1
Maximum	71	21	Diff_Pst_PreStand_D		1	NA		2
Average	10.19	8.87	Diff_Pre_Supine_StandS			K		3
Minimum	1	1	Diff_Pre_Supine_StandD		1	Ca	1	1
			Diff_Pre_Supine_StandD			Art_S		1
Parameters/rule			Pulse_Stand_Pre		3	Ven_S		
Maximum	3	6	Pulse_Supine_Post		2	SEX	1	
Average	1.44	2.6	Pulse_Stand_Post	2	1	diagnosis	12	5
			Age_at_Death			Age_Begin		1
<i>Trail set 8</i>								
Number of rules	25	18	Age_at_Death			diagnosis		6
Type of rules	Approximate	Absolute	Age_Begin		2	Off.Target		5
			Art_S			SEX		2
Cases described			Ca	2	4	Ven_S	2	
Maximum	20	17	d.Art_S	11		Wt_Loss		
Average	4.24	7.72	d.BFR		4			
Minimum	1	2	d.K	6	5			
			d.NA	19	6			
Parameters/rule			d.TBV	1				
Maximum	3	5	d.TD		6			
Average	1.64	2.22	d.Ven_S					

The results of the one-out-of n ($=114$) and k ($=10$) fold cross-validation for the 114 patient data set with transformed parameters are shown in Table 4. The transformation of parameters resulted in increase in the accuracy to 84.21% and 75.97% for one-out-of n and 10-fold cross validation approaches, respectively, over the initial untransformed parameter data mining.

The results produced by the DT and RS algorithms are compared in Table 5 (w.r.t. rules generated, parameters considered and cases observed) for Test set 4. Parameters such as Diff_Pst_PreSupine_D, Diff_Pst-PreStand_S, Pulse_Stand_Post, BFR, Calcium, and diagnosis were selected by DT and RS algorithm, indicating their importance in the dialysis treatment. Test set 8 considered only the discrete values of the chemicals in the dialysis solution. The results produced by the DT and RS algorithms for Test set 8 are compared in Table 5. Both DT and RS algorithms selected Calcium, sodium, and potassium, indicating their significance.

RS rules generated from Test set 4

Rule 1: IF (Diagnosis in {poly_kid_dis, ht_w_vas, con_neph_syn, ren_art_occ, weg_vas, CRF_w_c, ren_vas_dis, sys_scler, CRF}) THEN (Survival_Length = Below_med); [20.37%]

- Rule 2: IF (Diff_Pst_PreSupine_D \geq -11.7526) AND (TD in [192.195, 243.576]) AND (NA \geq 139.37) AND (K \geq 1.13712) AND (Art_S \geq -212.444) AND (Diagnosis in {diabetes, f_seg_glom, chr_p_re_neph, ht, lupus_e, neph_hero, na}) THEN (Survival_Length = Below_med); [29.63%]
- Rule 3: IF (Diff_Pst_PreStand_D \geq -11.4416) AND (Pulse_Stand_Pre \geq 58.7623) AND (TD \geq 243.576) AND (Ca \geq 2.99493) THEN (Survival_Length = Below_med); [24.07%]
- Rule 4: IF (Pulse_Stand_Pre \geq 58.7623) AND (Pulse_Supine_Post $<$ 64.2312) AND (K \geq 2.04805) THEN (Survival_Length = Below_med); [18.52%]
- Rule 5: IF (Diff_Pst_PreSupine_S $<$ 11.8505) AND (Diff_Pst_PreStand_S \geq -7.65919) AND (BFR $<$ 356.347) AND (Age_Begin \geq 40.8333) THEN (Survival_Length = Below_med); [18.52%]
- Rule 6: IF (Diff_Pst_PreStand_S \geq -23.5405) AND (TBV \geq 84.0784) AND (BFR \geq 404.087) THEN (Survival_Length = Below_med); [7.41%]
- Rule 7: IF (Off_Target $<$ 0.645559) AND (Pulse_Stand_Pre \geq 93.2237) THEN (Survival_Length = Below_med); [7.41%]
- Rule 8: IF (Diff_Pst_PreSupine_D in [-18.0055, -16.9068]) THEN (Survival_Length = Below_med); [1.85%]
- Rule 9: (Diagnosis in {glom, sol_kid_acq, poly_vas, poly_kid}) THEN (Survival_Length = Above_med); [16.00%]
- Rule 10: IF (Pulse_Supine_Post $<$ 79.6974) AND (TD \geq 194.428) AND (NA in [138.344, 139.861]) THEN (Survival_Length = Above_med); [42.00%]
- Rule 11: IF (TD \geq 192.195) AND (BFR $<$ 380.2) AND (K $<$ 1.70767) THEN (Survival_Length = Above_med); [16.00%]
- Rule 12: IF (DFR \geq 703.805) AND (Diagnosis in {ht, f_seg_glom}) THEN (Survival_Length = Above_med); [18.00%]
- Rule 13: IF (Diff_Pst_PreSupine_D $<$ -0.655757) AND (TD $<$ 192.195) THEN (Survival_Length = Above_med); [18.00%]
- Rule 14: IF (Off_Target $<$ 0.516172) AND (Pulse_Stand_Post $<$ 58.8233) AND (Diagnosis = ht) THEN (Survival_Length = Above_med); [16.00%]
- Rule 15: IF (Diff_Pre_Supine_StandD $<$ -12.8486) THEN (Survival_Length = Above_med); [2.00%]

DT rules generated from Test set 4

- Rule 1: IF (Diff_Pst_PreSupine_D \leq -12.23415) THEN (Survival_Length = Above_med) [85.7%]
- Rule 2: IF (Pulse_Stand_Post \leq 64.48718) AND (SEX = F) AND (Diagnosis = ht) THEN (Survival_Length = Above_med) [85.7%]
- Rule 3: IF (Diagnosis = glom) THEN (Survival_Length = Above_med) [83.3%]
- Rule 4: IF (Diagnosis = sol_kid_acq) THEN (Survival_Length = Above_med) [75.0%]
- Rule 5: IF (Diagnosis = poly_kid) THEN (Survival_Length = Above_med) [66.7%]
- Rule 6: IF (Diagnosis = poly_vas) THEN (Survival_Length = Above_med) [66.7%]

- Rule 7: IF (Diff_Pst_PreStand_S \leq -11.54973) THEN (Survival_Length = Above_med) [62.2%]
 Rule 8: IF (BFR \leq 252.4862) THEN (Survival_Length = Below_med) [85.7%]
 Rule 9: IF (Pulse_Stand_Post $>$ 64.48718) AND (Diagnosis = ht) THEN (Survival_Length = Below_med) [85.7%]
 Rule 10: IF (Diff_Pst_PreSupine_D $>$ - 12.23415) AND (Ca \leq 3.197368) AND (Diagnosis = diabetes) THEN (Survival_Length = Below_med) [81.8%]
 Rule 11: IF (Diff_Pst_PreSupine_D $>$ - 12.23415) AND (Diagnosis = f_seg_glom) THEN (Survival_Length = Below_med) [66.7%]
 Rule 12: IF (Diff_Pst_PreSupine_D $>$ - 12.23415) AND (Diagnosis = chr_p_re_neph) THEN (Survival_Length = Below_med) [66.7%]
 Rule 13: IF (Diagnosis = poly_kid_dis) THEN (Survival_Length = Below_med) [66.7%]
 Rule 14: IF (Diagnosis = ren_vas_dis) THEN (Survival_Length = Below_med) [66.7%]
 Rule 15: IF (Diagnosis = CiRF) THEN (Survival_Length = Below_med) [66.7%]
 Rule 16: IF (Diff_Pst_PreSupine_D $>$ - 12.23415) THEN (Survival_Length = Below_med) [60.3%]

5.3. Predictions

Survival can be predicted for patients with at least 15–20 visits, i.e., based on about a month of dialysis data. The results from the DT algorithm for test set of nine previously unseen cases are shown in Table 6. For six out of nine patients (66.67% of all test cases) the prediction was made by the majority of classifiers. There were four borderline cases where decisions (votes) resulted in a tie and the assigned decision was “Below-median”. After analyzing these cases it was observed that they (two out of four cases) in fact were nearly approaching the 3-year survival and there was a good chance of above the median survival. For cases where the survival time cannot be predicted with high accuracy, imply that a transplant in immediate future would be necessary.

The prediction made by the knowledge generated with the RS algorithm (simple voting and the weighted voting scheme) for the nine test cases are shown in Table 6. For five out of nine patients (56% of the test cases) the outcome was predicted by the majority of classifiers. Note that the outcome for patients 2 and 9 was predicted by all classifiers. This implies that these cases were “*most invariant*”. Thus the ranges of parameters values corresponding these patient forms a “*signature*” [18]. The commonality among patients (Table 6) was analyzed from the results generated by DT and RS algorithms. Thus for the patients 2, 4, 5, 9 the majority of classifiers predicted the same outcome irrespective of the type of the data mining algorithm used. For patients 6 and 7 erroneous predictions have been generated.

The knowledge generated by the DT algorithm produced higher classification accuracy, however, a limited numbers of parameters were included in the rules (Fig. 3). The RS algorithm yielded lower classification accuracy, produced more exact rules including more parameters. There is a tradeoff (Fig. 3) between overall prediction accuracy (higher for DT) and individual classification accuracy (higher for RS). There is a need for larger data sets to obtain more insights. As additional new test cases become available the confidence of the classifiers and decision-making model will increase.

Table 6
Prediction results and overlap analysis

Predictions: DT algorithm					
Case No.	Above_med	Below_med	Final_Pred	Actual_Pred	Result
1	5	3	Above_med	Above_med	Match
2	5	3	Above_med	Above_med	Match
3	4	4	Below_med	Above_med	Fail
4	4	4	Below_med	Below_med	Match
5	4	4	Below_med	Below_med	Match
6	4	4	Below_med	Above_med	Fail
7	2	6	Below_med	Above_med	Fail
8	1	7	Below_med	Below_med	Match
9	1	7	Below_med	Below_med	Match
Predictions: RS algorithm					
Case No.	Above_med	Below_med	Final_Pred	Actual_Pred	Result
1	0	2	Below_med	Above_med	Fail
2	8	0	Above_med	Above_med	Match
3	4	3	Above_med	Above_med	Match
4	2	2	Below_med	Below_med	Match
5	1	4	Below_med	Below_med	Match
6	1	4	Below_med	Above_med	Fail
7	1	2	Below_med	Above_med	Fail
8	3	0	Above_med	Below_med	Fail
9	0	8	Below_med	Below_med	Match
Overlap analysis					
Prediction		DT	RS Exact	RS Approx.	Overlap
Correct	Above_med	1, 2	2, 3	2	2
	Below_med	4, 5, 8, 9	4, 5, 9	4, 5, 9	4, 5, 9
Wrong	Above_med	3, 6, 7	1, 6, 7	1, 3, 6, 7	6, 7
	Below_med	—	8	8	—

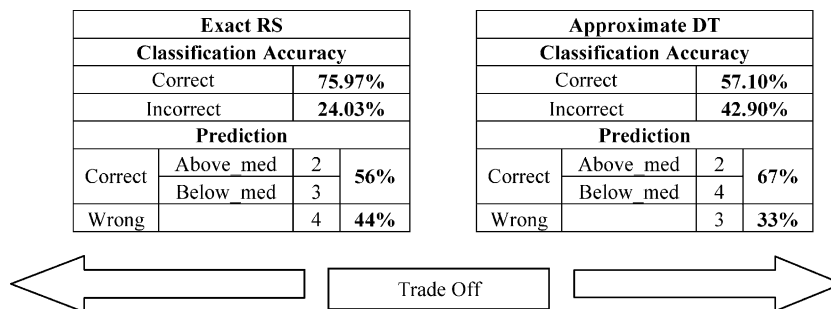


Fig. 3. Decision-maker's dilemma.

Table 7
Significant dialysis parameters

	Algorithm		Combined
	DT	RS	
Parameters	Diagnosis	Diagnosis	Diagnosis
	Arterial pressure	Total dialysis time	Total dialysis time
	Sodium level	Potassium level and deviation from target weight	Sodium level
	Post-dialysis pulse rate supine	Sodium level and blood flow rate	Arterial pressure
	Difference between post and pre supine (LL)	Calcium level	Deviation from target weight

6. Medical significance

After analyzing the decision rules generated by RS and DT algorithms, significant parameters were identified based on the high strength rules (Table 7).

The type of initial diagnosis played the most important role in the rules generated by both data-mining algorithms. This is supported by the fact that the condition of a patient before dialysis greatly affects the patient survival. The two main causes of kidney failure are diabetes (initiation factor) [1] and high blood pressure (progression factor) [3,19]. Studies show that patients (within the age group of 55–64) with diabetes have shorter survival rates [3,20]. On the positive side, survival time increases with better control over blood pressure and blood glucose [21].

The arterial blood pressure is responsible for driving blood forward into the tissues throughout the cardiac cycle. The patient's arterial pressure is measured as the systolic and diastolic pressures, the cardiac frequency, and the hemoglobin concentration. If the ultrafiltration rate is too high, the patient may collapse due to a sudden increase of extra-cellular fluid. By monitoring the person's arterial pressure, the clinician can determine the patient's plasma volume. If the arterial blood pressure drops or increases above a certain threshold, adequate blood flow is not achieved and can be controlled by adjusting the bulk blood flow rate parameter.

The inability of the kidneys to adjust sodium excretion to balance the changes in sodium consumption results into sodium imbalances. Side effects include elevated blood pressure, congestive heart failure, and hypotension. If the sodium level cannot be controlled, this most likely means the person has a heart condition and/or high blood pressure, which contribute to the decrease in survival rate.

The less time it takes for dialysis session to be completed, the smaller the amount of urea being removed. The efficiency of the removal of uremic toxins is usually assessed by the urea clearance fractional rate parameter, Kt/V [7]. This parameter expresses the fraction of blood volume that is cleared per unit time. The clearance, K , is a function of the blood bulk flow in each dialysis session (QB). The urea distribution volume is directly proportional to the body weight (BW) [7]. The dialysis time (t) influences the clearance fractional rate (Kt/V). The evaluation of clearance rate (Kt/V) is based on QB, BW, and t values [7]. These three factors result in an improved outcome and influence

the quality of the dialysis treatment for a patient [7,22–24]. The last parameter of significance is the deviation from target weight. The clinician estimates a patient should lose around 1–2 kg with each dialysis session. If a patient loses less than that amount, the quantity of urea cleared decreases indicating deterioration in the treatment quality and lower survival rate. As mentioned above, the clearance rate (Kt/V) is managed by monitoring body weight (BW) [7].

Weight, sodium level, total dialysis time, arterial blood pressure, and diagnosis are all related to each other as inferred from the above discussion. Problems with sodium levels, arterial blood pressure, total dialysis time, and weight can indicate the diagnosis of the patient, for example, heart disease. Controlling of these parameters for patient with heart disease may not be trouble-free, thus leading to a decrease in survival rate.

7. Conclusions

7.1. Research results

The most significant result obtained from this research was to demonstrate that data mining, data transformation, data partitioning, and decision-making algorithms are useful for survival prediction of dialysis patients. The potential for making accurate decisions for individual patients is enormous and the classification accuracy is high enough (above 75–85%) to warrant use of additional resources and conduct further research. Data transformation increased the classification accuracy by approximately 11%. Analyzing and comparing the data mining rule sets produced a list of significant parameters, such as the diagnosis, total dialysis time, potassium, calcium and sodium levels, deviation from target weight, arterial pressure, post-dialysis pulse rate supine, difference between post- and pre-supine. The medical relevance of the significant parameters was established. The decision-tree algorithm produced correct predictions 67% of the time for the test data set. The RS algorithm produced correct predictions 56% of the time using a simple as well as the weighted voting scheme on the test data set.

7.2. Benefits and applications of the research results

Clinical studies: Accurate prediction of outcomes is of great interests to clinical studies where patients with particular characteristics are sought. The approach presented in the paper reduces the cost and effort of selecting patients for clinical studies. Patients can be chosen based on the prediction results and the most significant parameters discovered.

Treatment selection: Depending on the predicted outcome, a more suitable treatment protocol can be selected for an individual patient. Therefore, the necessary interventions can take place at an appropriate time of the treatment process. It is important to note here that significant parameters discovered in this research may be not general enough due to limited size of the database.

“Model patient” selection: The patients for whom the outcomes are correctly predicted by a large number of classifiers are of interest. Such patients can be considered “model patients”.

More effective data collection process: This research shows that the value of the dialysis data could be increased if the data was systematically collected. The original volume of data collected in this project was in excess of 500 Mbytes, however, because the data was collected for reasons other

than data mining, only a limited subset was usable for knowledge extraction. Through this process it was discovered that the number of parameters in the data could be reduced, however, the data collection process needs to be more systematic.

7.3. Future research

The data set used for this research was rather small. A larger set could provide more meaningful results. Another issue is the incompleteness of data. In this research, several parameters, e.g., demographics, were not considered in the analysis. The parameters were difficult to convert in a form acceptable for data mining and were therefore discarded. The dialysis machine was only able to record 15–20 readings throughout each session and a final session reading this was due to the limited memory of the machine. If these several missing readings per session would be added to the data set, they could provide additional information.

The data used in this research was compiled from numerous data tables. Therefore, the patient parameters were not considered over time (temporal data). Temporal data could contain valuable information about the patient health [25]. It is important to obtain missing information regarding the stage of dialysis of each patient. In the future data mining could be performed on the data collected for individual sessions. A controlled population of patients could be used. Predicting a “short-term performance” parameter (e.g., urea clearance fraction (Kt/V)) could be investigated.

Acknowledgements

Our special thanks to Leah Bontrager, Leah Bruxvoort, and Michelle Reyes for preparation of different versions of data sets and Amy Conroy for medical significance related inputs.

References

- [1] US Renal Data System, USRDS 2002 Annual Data Report: Atlas of End-Stage Renal Disease in the United States, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, Accessed on 2002, December 03, Website: <http://www.usrds.org/atlas.htm>.
- [2] J. Cooper, US incidence of kidney failure is the highest in the world, The Medical Reporter, Accessed on 2002, April 30. Website: <http://medicalreporter.health.org/tmr0799/kidney.html>.
- [3] K/DOQI Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification, National Kidney Foundation, Accessed on 2002, December 03. Website: <http://www.kidney.org/professionals/doqi/kdoqi/toc.htm>.
- [4] National Institute of Diabetes & Digestive & Kidney Diseases, National Kidney and Urologic Diseases Information Clearinghouse, Your Kidneys and How They Work, NIH Publication No. 02-4241. February 2002, Website: www.niddk.nih.gov/health/kidney/pubs/yourkids/index.htm.
- [5] L. Sherwood, Human Physiology: From Cells to Systems, 3rd Edition, Wadsworth Publishing Company, Albany, NY, 1993.
- [6] R. W. Hamilton, Principles of dialysis: diffusion, convection, and dialysis machines, in: W.L. Henrich, W.M. Bennet (Eds.), Atlas of Diseases of the Kidney, Vol. 5, 1999, Website: <http://www.kidneyatlas.org/book5/adk5-01.ccc.QXD.pdf> (On-line edition: ISN Informatics Commission and NKF cyberNephrology).
- [7] R. Bellazzi, C. Larizza, P. Magni, R. Bellazzi, S. Cetta, Intelligent Data Analysis Techniques for Quality Assessment of Hemodialysis Services, Accessed on 2002, April 30, Website: <http://magix.fri.uni-lj.si/idamap2001/papers/bellazzi.pdf>.

- [8] The United States Renal Data Systems, Accessed on 2002, April 30, Website: www.usrds.org.
- [9] PAKDD Workshop, Toward the Foundation of Data Mining, Taipei, Taiwan, Accessed on 2002, April 30. Website: www.mathcs.sjsu.edu/faculty/tylin/pakdd_workshop.html.
- [10] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, MA, 1997.
- [11] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer, Boston, MA, 1991.
- [12] M. Stone, Cross-validatory choice and assessment of statistical predictions, *J. R. Stat. Soc.* 36 (1974) 111–147.
- [13] A. Kusiak, Feature transformation methods in data mining, *IEEE Trans. Electron. Packag. Manuf.* 24 (2001) 214–221.
- [14] R. Cattral, F. Oppacher, D. Deugo, Supervised and unsupervised data mining with an evolutionary algorithm, in: *Proceedings of the 2001 Congress on Evolutionary Computation*, IEEE Press, Piscataway, New Jersey, 2001, pp. 767–776.
- [15] H. Vafaie, K. De Jong, Feature space transformation using genetic algorithms, *IEEE Intell. Systems* 13 (1998) 57–65.
- [16] N. Lesh, M.J. Zaki, M. Ogihara, Scalable feature mining for sequential data, *IEEE Intell. Systems* 15 (2000) 48–55.
- [17] R. Quinlan, *C 4.5 Programs for ml*, Morgan Kaufmann, San Mateo, CA, 1992.
- [18] A. Kusiak, A data mining approach for generation of control signatures, *ASME Trans. J. Manufactur. Sci. Eng.* 124 (2002) 923–926.
- [19] The Kidney Foundation of Canada. Accessed on 2002, April 30. Website: www.kidney.ca/per-e.htm.
- [20] C. Byrne, P. Vernon, J.J. Cohen, Effect of age and diagnosis on survival of older patients beginning chronic dialysis, *JAMA* 271 (1994) 34–36.
- [21] M. Schomig, A. Eisenhardt, E. Ritz, Controversy on optimal blood pressure on haemodialysis: normotensive blood pressure values are essential for survival, *Nephrol. Dial. Transplant.* 16 (2001) 469–474.
- [22] W.E. Bloembergen, D.C. Stannard, F.K. Port, R.A. Wolfe, J.A. Pugh, C.A. Jones, J.W. Graer, T.A. Golper, P.J. Held, Relationship of dose of hemodialysis and cause-specific mortality, *Kidney Int.* 50 (1996) 557–565.
- [23] S.J. Davies, L. Phillips, A.M. Griffiths, P.F. Naish, G.I. Russell, Analysis of the effects of increasing delivered dialysis treatment to malnourished peritoneal dialysis patients, *Kidney Int.* 57 (2000) 1743–1754.
- [24] T.F. Parker, L. Husni, W. Huang, N. Lew, E.G. Lowrie, Survival of hemodialysis patients in the United States is improved with a greater quantity of dialysis, *Am. J. Kidney Dis.* 23 (1994) 670–680.
- [25] S. Hedberg, Stanford University's AI in medicine: still cutting the edge, *IEEE Intell. Systems* 13 (1998) 74–76.

Andrew Kusiak is a Professor of Industrial Engineering at the University of Iowa, Iowa City. He is interested in theory and applications of computational intelligence, data mining, and optimization in healthcare, pharma, product development, and manufacturing. He has published research papers in journals sponsored by AAAI, IEEE, IIE, INFORMS, ESOR, IFIP, IFAC, IPE, ISPE, and SME. He speaks frequently on international meetings, conducts professional seminars, and consults for industrial corporations. He serves on the editorial boards of 16 journals, and edits book series. He is the Editor-in-Chief of the *Journal of Intelligent Manufacturing*.

Bradley S. Dixon is an Associate Professor of Medicine in the Division of Nephrology at the University of Iowa Roy J. and Lucille A. Carver College of Medicine and staff physician at the Iowa City Veterans Affairs Medical Center. His clinical research is devoted to improving health outcomes for people with chronic kidney disease particularly in the areas of preventing vascular access failure and cardiovascular disease. Basic research in his lab is focused on understanding signal transduction pathways that regulate vascular smooth muscle cell proliferation, which contributes to the neointimal hyperplasia leading to vascular access failure and cardiovascular disease.

Shital Shah is a Ph.D. student in Industrial Engineering at The University of Iowa, Iowa City. He completed MS in IE at The University of Alabama, Tuscaloosa and BE in production engineering at VJTI, Bombay, India. His area of interest are in computational intelligence, data mining, informatics, operations research, simulation, and decision support systems. He is in the process of publishing various data mining and informatics related research papers. He is the member of ALPHA-PI-MU and IIE.