

Data Mining in Predicting Survival of Kidney Dialysis Patients - Invariant object approach

Shital Shah*, Andrew Kusiak*, and Bradley Dixon**

* Intelligent Systems Laboratory, 3221 SC, **University of Iowa Hospital and Clinics, E300D GH,
The University of Iowa, Iowa City, IA 52242-1527

ABSTRACT

The number of patients on hemodialysis due to end stage kidney disease is increasing. The median survival for these patients is only about 3 years and the cost of providing care is high. Finding ways to improve patient outcomes and reduce the cost of dialysis is a challenging task. Dialysis care is complex and multiple factors may influence patient survival. More than 50 parameters may be monitored while providing a kidney dialysis treatment. Understanding the collective role of these parameters in determining outcomes for an individual patient and administering individualized treatments is of importance. Individual patient survival may depend on a complex interrelationship between multiple demographic and clinical variables, medications, and medical interventions. In this research, a data mining approach is used to elicit knowledge about the interaction between these variables and patient survival. Two different data mining algorithms are employed for extracting knowledge in the form of decision rules. Data mining is performed on the individual visits of the "most invariant" patients as they form "signatures" for their decision categories. The concepts introduced in this research have been applied and tested using a data collected at four dialysis sites. The computational results are reported.

KEYWORDS: Data mining, Decision-making, Kidney, Dialysis, Hemodialysis, Invariant objects, Signatures.

1. INTRODUCTION

Approximately 370,000 Americans suffer from end stage kidney disease and require some form of renal replacement therapy (either dialysis or kidney transplantation) to sustain life¹. Statistics from the United States Renal Data Systems in the year 2000 reported that nearly 250,000 Americans required regular hemodialysis¹. The annual cost for treating kidney disease is over \$12 billion¹. The number of Americans with end stage kidney failure is steadily growing. According to the National Kidney Foundation, around 40 million Americans are at increased risk for developing chronic kidney disease². United States leads the world in the number of new cases of kidney failure per million of population^{1 3}. Chronic kidney disease occurs when the kidneys are functioning at less than 50% of normal capacity². End-stage kidney disease occurs when the kidneys are working at less than 10-15% of normal capacity⁴. At this stage, either transplantation or repetitive kidney dialysis becomes necessary for survival. Hemodialysis is typically carried out in a clinic setting and accounts for more than 90% of the dialysis population, with peritoneal dialysis accounting for the remaining 10 %.

The primary role of the kidneys is to remove metabolic waste products and to maintain water and electrolyte balance. The kidneys receive about 25% of the total blood pumped by the heart and filter nearly 187 liters of liquid per day. About 1% of the original liquid filtrate ultimately appears in the final urine as waste products and extra water. The kidneys reabsorb the remaining 99% of initial filtrate into the blood stream to prevent dehydration. 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 stability of the extra-cellular fluid (ECF) volume and electrolyte homeostasis by adjusting excretion of water and electrolytes to balance changes in intake⁵. If any of these nutrients are deficient, the kidneys can conserve these nutrients until they are replenished through ingestion. 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 components⁵ from the body. 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 death⁵.

The excretory function of the kidney can be replaced by dialysis. In hemodialysis, blood is removed from the body and passed through a large number of very small porous tubes that are continuously bathed in a fluid known as dialysate. The dialysate has a defined electrolyte and chemical composition. Toxic substances in the blood diffuse through the dialysis membrane and are removed in the dialysate. The composition of the dialysate sets the lower limit for diffusion of electrolytes and glucose from the blood to prevent excessive loss of these substances. Excess water in the body can be removed by setting the pressure across the dialysis membrane to remove fluid (ultrafiltration). The purified and concentrated blood is then returned to the patient. A typical hemodialysis session lasts 4 hours and is repeated 3 times per week. The length and frequency of dialysis as well as the composition of the dialysate and the amount of fluid to be removed are some of the many variables prescribed by the physician. This is called the dialysis prescription and along with medications to replace the endocrine function of the kidney and to control other side effects such as high blood pressure, constitutes the overall prescription needed to manage people with kidney failure. Although dialysis is life saving for a person with terminal kidney failure, survival is still markedly reduced compared to healthy people of a similar age. Understanding what factors are predictive of survival in a given dialysis patient may allow targeted intervention for high-risk patients and suggest areas where improvements in the dialysis prescription or further research might improve survival⁶.

Hemodialysis patients have a large volume of medical data collected over time. The large volume of data makes it difficult for clinicians to spot patterns in the data that may be related to the outcome for an individual patient or group of patients. Data mining provides algorithms and tools for identifying valid, novel, potentially useful, and ultimately understandable patterns from data^{7 8}. The discovered patterns may represent valuable knowledge that could lead to medical discoveries, e.g., combinations of parameter values that lead to a longer survival time. The collection of settings of the dialysis machine and medications by the patient is called a dialysis treatment protocol. The dialysis protocol depends on the patient characteristics (data) e.g. age, disease, etc. The relationship between the dialysis parameters, patient data, and the dialysis outcomes are not well understood. Thus in this paper a data mining approach was applied to individual visit information of the most invariant patients. This was done to determine which factors/features are predictive of an individual patient surviving beyond the median survival time.

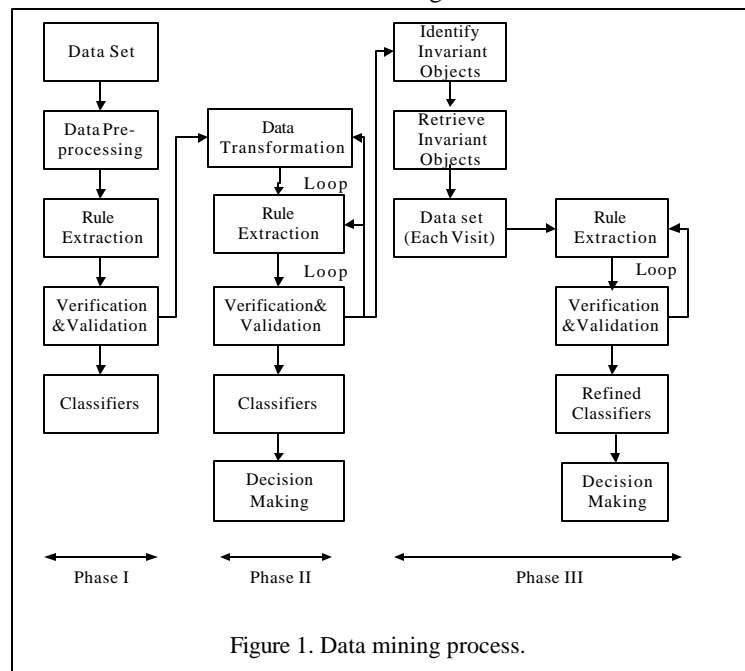


Figure 1. Data mining process.

2. DATA MINING PROCESS

The data from distributed databases was merged into a single data set that was preprocessed by computing averages, ignoring records with missing data, filling in missing data, etc. Initial results are obtained without the inclusion of problem dependent transformation and/or domain knowledge. This process could lead to discovery of previously unknown patterns from unrelated features as per the domain experts. Results were verified with the domain expert and validated with new sample test cases. These activities form the Phase I of the research (Figure 1).

In Phase II, data transformation⁹ was carried out. It is done by using domain knowledge, combining features, and using statistical methods. A new set of rules was extracted that was verified and validated. Thus depending upon the results

one of the three steps is performed: data transformation, knowledge extraction, and generation of invariant objects (see Figure 1). In Phase III (Figure 1) "Most Invariant" patients are identified based on previous knowledge and Phase II results. The most invariant patients can be based on the average values of all features, i.e., an aggregate data set. Each patient's visit can be used as a stand-alone record and knowledge extraction is performed on the individual visit data sets. The motivation behind this approach is *Reverse Engineering*. In this reverse engineering approach a data set containing individual visits for a relatively small number of patients is mined. The extracted knowledge is used for predictions for a large number of patients on either individual visits or an aggregate basis. To improve classification accuracy, insignificant parameters and patient data could be deleted from the data set. An approach similar to Phase II is followed for the data set with individual visits. The classifiers developed from this most invariant representative data set refine the initial classifiers developed from previous phases of data mining. The constructed classifiers contain more relevant features and are more accurate.

3. DATA PREPARATION

The University of Iowa and its four satellite locations have been collecting dialysis data from several sources, in some cases for nearly 15 years. These data include information on the dialysis prescription, data electronically collected during each dialysis treatment, laboratory tests, pharmacy records, patient diagnosis and demographic data. Before each session, the patient was weighed and her/his blood pressure (systolic and diastolic) registered while sitting (supine pressure), and when possible, while standing. The weight and blood pressure measurements are repeated at the end of the session. The levels of sodium, bicarbonate, potassium, calcium, and glucose in the dialysis solution are recorded. Total time for the dialysis session, blood flow rate, total volume of blood processed, dialysis flow rate, and the overall average pressures at the arterial and venial side of the blood pump were another set of collected values. A set of measurements was collected by the dialysis machine every twenty minutes or on request. This set includes systemic blood pressure, pulse, blood flow rate, arterial and venial blood pump pressures, trans-membrane pressure, and the rate of ultrafiltration. To reduce data noise, averages were computed over the fifteen readings taken by the machine during the dialysis session.

The demographic and outcomes data set contains the patient's date of birth, gender, and race; the date(s) of death, kidney transplant, and transfer into or out of the dialysis center. The final portion contains the diagnosis codes for the primary and secondary diagnoses. Differences between each patient's average post and pre systemic blood pressures were calculated for all four combinations of systolic and diastolic pressures and supine and standing positions. The pulse pressures (determined by the difference between the systolic and diastolic blood pressures) were calculated for both pre and post conditions for both supine and standing positions. Differences between the supine and standing pressures were also calculated for both systolic and diastolic blood pressure and for the pre and post dialysis conditions. Some new features were added to the data set by using the concept of data transformation⁶.

Merging files from four outreach centers created a data set for 188 patients. Averages were computed for each patient for all variables to form a single representative record (aggregate data set). Initial data mining focused on a selected group of long-term dialysis patients with at least fifteen or more visits. In the next phase of analysis most invariant objects were identified (for each decision categories) and records of individual visits were retrieved. Data cleaning and transformations similar to the aggregate data set was performed so as to form a common basis for comparison.

4. DATA MINING

Three decision categories were developed for data mining purposes. The category "Above Median" includes patients who have either survived more than three years on dialysis and are still alive or deceased (Figure 2)⁶. The second category "Below Median" consists of patients who survived less than three years on dialysis (Figure 2). The third category "Indeterminate" includes patients who are alive and have not yet completed three years on dialysis. The group of patients in the "Not determined" category (Figure 2) could land in either categories of above or below median and were excluded from data sets for data mining.

4.1. Preliminary data mining

Initially the aggregate "as-is" data set was mined, meaning no changes were made to the original preprocessed data set. Thereafter, data mining was performed on transformed data with the decision variable with four discrete values and two discrete values⁶. Special consideration (for inclusion in data mining procedure) was given for the patients who received transplants in order to be classified into the one of the categories described in Figure 2 Rough set theory (RST) and decision tree (DT) algorithms were used for data mining.

	Alive	Deceased
Survived three plus years	Above median	
Survived less than three years	Not determined	Below median

Figure 2. Decision value formulation.

Sixteen different classifiers were developed using eight sub-data sets from master aggregate data set and two data mining algorithms⁶. A decision-making algorithm was developed for evaluating the predictions from sixteen classifiers. Some patients in the " Not determined " category (Figure 2) actually crossed in the above or below median categories at the end of the study and were used for verification and validation of the sixteen classifiers and decision-making algorithm was tested for the predictability of data mining tools. The rough set theory (RST) algorithm generated rules with a higher level of confidence, exhibited higher reliability and had overall lower misclassification. The decision tree (DT) algorithm generated rules containing fewer features than the rules produced by the RST 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. Certain significant features (such as Diagnosis, Total dialysis time, Arterial pressure, Potassium level, Deviation from target weight, Calcium level, Blood flow rate, Post-dialysis pulse rate supine) and their ranges were identified by data mining procedures to be incorporated into the individualized dialysis treatment protocols.

4.2. Data Mining of Invariant Objects

From the previous study Case 1 (decision = Above Median) and Case 2 (decision = Below Median) predicted the same outcome for each test partial data set. This implied that these cases were "*Most Invariant*" in nature. Similarly Cases 3, 4, and 5 (decision = Above Median) and Cases 6, 7, and 8 (decision = Below Median) were of interest as they too fitted into the category of most invariant patients. Thus these most invariant patients formed "*Signatures*"¹⁰ for that outcome and is desired that other patient conform to the set of feature values of this signature. Each case had a large number of visits and thus each visit was used as a stand-alone record forming individual visit data set for data mining.

Data mining was conducted using a rough set algorithm. In the RST lower and upper approximations of the concept are computed¹¹. As a result, there are two types of decision rules, *certain* or *approximate*. *Certain* rules are those that are induced from the lower approximation, where lower approximation refers to the set of observations that can all be classified into the given concept. The *approximate* rules are those induced from the upper approximation, where upper approximation refers to the set of observations that can be possibly classified into this concept¹¹. Classification accuracy can then be 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 set¹¹. DT was chosen as an alternative data-mining tool to explore the dialysis data. Decision-tree algorithm creates rules based on decision trees or sets of if-then statements to maximize interpretability¹². Through the use of recursive partitioning of data, the decision-tree algorithm creates a classification-decision tree using depth-first strategy. The decision-tree algorithm considers all of the plausible tests that can split the data set based on best information gain.

There were 8 patients each having around 80 to 400 visits. Thus a total of more than 2000 records with over fifty features were considered for data mining. Certain significant features from previous study such as Diagnosis, Total Dialysis time, and Target weight were removed from further analysis. These features were already found to be significant for inclusion into dialysis treatment protocol. Exclusion of some features allowed a "drill-down" approach for the next level of significant features to be incorporated into the protocol. Also the small number of patients used for analysis didn't offer a uniform sample covering all the possible feature values of the excluded features.

5. COMPUTATIONAL RESULTS

The classification accuracy for individual visit data set (each visit information was considered as an independent patient record) (for both RST and DT algorithms) is noticeably higher with a marked improvement over the aggregate data set (Table 1). These improvements can be attributed to the data mining of most invariant cases as these cases exhibit certain characteristics, which are unique to each decision category. Most invariant cases act as a representative sample for their decision category and the feature values cover most of the patients. Improved rule set combined with high classification accuracy yielded higher prediction accuracy (Table 3).

In Phase II, aggregate classifiers or rule set were formed for both above median and below median decisions and these rule sets were used for prediction of new cases/patients. Each new case either fitted the rules sets of the decision category without conflicts or had conflicts or did not fit at all. The first case is clear-cut decision in the respective category, while the second can be handled by voting schemes. The third can be handled based on the feature distances from either category rule set (Figure 3). The hypothetical case 1 has b11 distance from below median and a11 distance from above median (Figure 3). Thus depending on the distances, the cases can be handled as above or below median. The predictions for these hypothetical cases are in Table 2.

Table 1. Comparison of classification accuracy.

	Aggregate Data Set			Individual Visit Data Set		
	Correct	Incorrect	None	Correct	Incorrect	None
Below Median	52.83	47.17	0	97.17	0.60	2.24
Above Median	95.75	4.25	0	96.66	0.63	2.71
Average	75.97	24.03	0	97.78	0.55	1.66

Table 2. Predicated decisions for hypothetical cases.

Hypothetical Cases	Aggregate Data set Phase II	Individual visit Data set Phase III
	Decisions	Decisions
Case 1	Below Median	Below Median
Case 2	Above Median	Above Median
Case 3	Either way	Below Median
Case 4	Above Median	Above Median
Case 5	Below Median	Below Median

Data mining on data set with individual visits for most invariant cases in Phase III has improved the rule set quality. The improvement is measured in terms of enhancement in classification accuracy. The most invariant patients lay at the farthest ends in Figure 3 as per their decision categories and the gap between the ends is more in Phase III as compared to Phase II. The hypothetical cases are still at the same positions and their distances from the Phase III rule sets are reallocated. The increase in gap and the rule set with the farthest condition patients will provide better framework for decision-making and essentially improve the prediction accuracy. Hypothetical cases are reclassified as per Phase III rule set (Table 2).

Several iterations of data mining over a number of separate data sets will help to tighten the upper and lower bound (similar to optimization) of the rule sets. A pool of most invariant cases from different data sets is used to develop the optimal/ sub-optimal rule sets. These rule sets will exhibit the ideal conditions for both categories. Thus the improvement in the features values exhibited at optimal/sub-optimal below median will lead to improvement in survival time. The best operating condition will be at the feature values of above median rules of Phase III. Thus an individualized dialysis treatment protocol can be developed.

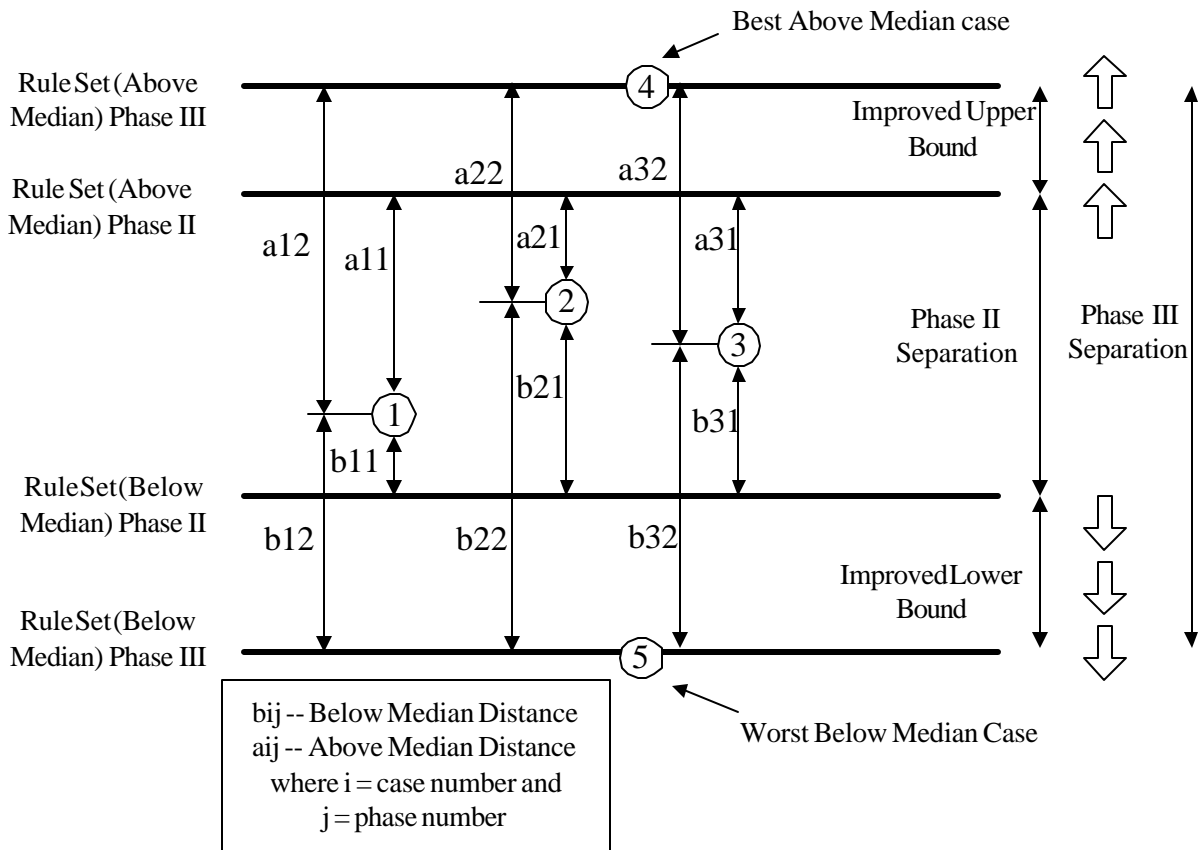


Figure 3. Optimization concept in data mining.

6. DECISION MAKING

Data mining in the final phase was carried out using an approach of partial individual visit data set mining⁶. The grouping of features for partial data sets was prepared, keeping in mind medical relevance between these features (e.g. dialysis chemical solution, weight, blood pressure, difference in blood pressure (i.e. pulse pressure), etc. Eleven different combinations were determined to form trial data sets. These eleven data sets were mined separately using rough set based and decision-tree based data mining algorithms. Each data subset produced two sets of rules (classifiers), one each from the two data mining algorithms. Thus in all there were twenty-two classifiers capable of predicting the outcomes for new patients. These classifiers were developed to perform multi-angle, highly reliable (parallel redundancy concept in reliability engineering), robust, accurate decisions/predictions. The classifiers can be combined to form a single classifier, which could be used for prediction of new patients or individual classifiers could come with their own prediction and these predictions, could be combined by using voting/weighted-voting schemes. There was considerable increase in the prediction accuracy of individual visit over the aggregate data set (Table 3).

Table 3. Prediction accuracy comparison.

Data set	Prediction Accuracy			
	Aggregate		Individual visit	
Algorithm	Exact RST	DT	Exact RST	DT
Correct	0.56	0.67	0.65	0.65
Wrong	0.44	0.33	0.35	0.35

7. MEDICAL SIGNIFICANCE

The significant features identified by data mining algorithms after Phase II and Phase III are as follows diagnosis, time on dialysis, deviation from target weight, blood pressures ranges for different patients, calcium and potassium levels in dialysis solution, total blood volume, blood flow rate, venial pressures.

The two main causes of kidney failure are diabetes (initiation factor)¹ and high blood pressure (progression factor)^{13 2}, which were also identified by the data mining algorithms. A study shows that patients within the age group of 55-64 (susceptibility factor) with diabetes have shorter survival rates^{14 2}. On the positive side, survival time increases with better control over blood pressure and blood glucose¹⁵. The efficiency of the removal of protein catabolism is usually assessed by the urea clearance fractional rate parameter, Kt/V ¹⁶. 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)¹⁶. The dialysis time (t) influences the clearance fractional rate (Kt/V). This clearance rate (Kt/V), is evaluated by monitoring QB , BW , and t ¹⁶. These three factors influence the quality of the dialysis treatment for a patient as per Bellazzi *et al.* (2002) and were verified by data mining algorithms. Also factors depending upon the Kt/V such as off-target weight loss, time of dialysis, and chemicals (calcium and potassium) of dialysis solution were identified as important features by data mining procedure.

8. CONCLUSIONS

The overall classification accuracy for all data mining algorithms was significantly higher using the individual visit data set over the aggregate data set (Table 1). The prediction accuracy of individual visit based rule sets increased over the aggregate based rule sets (Table 3). These improvements were in spite of excluding known significant features such as diagnosis, time for dialysis, target weight, etc. Thus the confidence in the individual data set classifiers is considerably higher. The significant features (identified collectively by data mining) that could be included in the dialysis treatment protocol are diagnosis, time on dialysis, deviation from target weight, blood pressures ranges for different patients, calcium and potassium levels in dialysis solution, total blood volume, blood flow rate, venial pressures. Some of these are known from previous work while others such as venous pressure or calcium in the bath need more investigation. Future work is needed to understand how they influence survival. The information obtained from these studies can be used for future clinical studies, treatment selections, refinement for data sets, and data collection protocols (relatively less number of features). For further research temporal data could be derived by data transformations as it may contain valuable information about the patient health¹⁷. Predicting a “short term performance” parameter (e.g., urea clearance fraction (Kt/V)) could also be investigated.

ACKNOWLEDGEMENT

Our special thanks to Katherine Ruth Coates, Sara Ann Dolny, Katherine Gaunt, and Katie Louise Kruse for preparation different versions of data sets.

REFERENCES

- [1] USRDS 2002, U.S. 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*, <http://www.usrds.org/atlas.htm>, Bethesda, MD, 2002.
- [2] K/DOQI, K/DOQI Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification, *National Kidney Foundation*, <http://www.kidney.org/professionals/doqi/kdoqi/toc.htm>, Bethesda, MD, 2002.
- [3] Cooper, J., U.S. incidence of kidney failure is the highest in the world, *The Medical Reporter*, <http://medicalreporter.health.org/tmr0799/kidney.html>, accessed April 30, 2002, 1999.
- [4] NIDDK, National Institute of Diabetes & Digestive & Kidney Diseases, National Kidney and Urologic Diseases Information Clearinghouse, *Your Kidneys and How They Work*,

S. Shah, A. Kusiak, and B. Dixon, Data Mining in Predicting Survival of Kidney Dialysis Patients, in *Proceedings of Photonics West - Bios 2003*, Bass, L.S. et al. (Eds), *Lasers in Surgery: Advanced Characterization, Therapeutics, and Systems XIII*, Vol. 4949, SPIE, Bellingham, WA, January 2003.

- www.niddk.nih.gov/health/kidney/pubs/yourkids/index.htm, NIH Publication No. 02-4241, February 2002, accessed April 30, 2002.
- [5] Sherwood, L., *Human Physiology: From Cells to Systems*, Third Edition, Wadsworth Publishing Company, Albany, NY, 1993.
 - [6] Kusiak, A., Dixon, B., and Shah, S., "Predicting Survival Time for Kidney Dialysis Patients: A Data Mining Approach", Intelligence System Laboratory, Working Paper ISL_SCS001_017, 2002.
 - [7] PAKDD Workshop, Toward the Foundation of Data Mining, www.mathcs.sjsu.edu/faculty/tylin/pakdd_workshop.html, Taipei, Taiwan, accessed April 30, 2002.
 - [8] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds., *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, MA, 1997.
 - [9] Kusiak, A., "Feature Transformation Methods in Data Mining", *IEEE Transactions on Electronics Packaging Manufacturing*, **Vol. 24**, No. 3, pp. 214 –221, 2001.
 - [10] Kusiak, A., "A Data Mining Approach for Generation of Control Signatures", *ASME Transactions: Journal of Manufacturing Science and Engineering*, **Vol. 124**, No. 4, pp. 923-926, 2002.
 - [11] Pawlak, Z., *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer, Boston, MA, 1991.
 - [12] Quinlan, R., *C 4.5 Programs for ml*, Morgan Kaufmann, San Mateo, CA, 1992.
 - [13] KFC_1, The Kidney Foundation of Canada, www.kidney.ca/per-e.htm, accessed April 30, 2002.
 - [14] Byrne, C., Vernon, P., Cohen, J., J., "Effect of Age and Diagnosis on Survival of Older Patients Beginning Chronic Dialysis", *JAMA*, **Vol. 271**, Issue.1, pp.34-36, 1994.
 - [15] Schomig, M., Eisenhardt, A., and Ritz, E., "Controversy on optimal blood pressure on haemodialysis: normotensive blood pressure values are essential for survival", *Nephro Dial Transplant*, **Vol. 16**, 2001, pp. 469-474, 2001.
 - [16] Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R., Cetta, S., *Intelligent Data Analysis Techniques for Quality Assessment of Hemodialysis Services*, <http://magix.fri.unilj.si/idamap2001/papers/bellazzi.pdf>, accessed April 30, 2002.
 - [17] Hedberg, S., "Stanford University's AI in medicine: Still cutting the edge", *IEEE Intelligent Systems*, **Vol.13**, Issue.1, pp.74 –76, 1998.