2007

Modeling of hemodialysis patient hemoglobin: a data mining exploration

Michael Francis Bries
University of Iowa

MODELING OF HEMODIALYSIS PATIENT HEMOGLOBIN:

A DATA MINING EXPLORATION

by

Michael Francis Bries

A thesis submitted in partial fulfillment
of the requirements for the Master of
Science degree in Industrial Engineering
in the Graduate College of
The University of Iowa

May 2007

Thesis Supervisor: Professor Andrew Kusiak

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

MASTER'S THESIS

This is to certify that the Master's thesis of

Michael Francis Bries

has been approved by the Examining Committee
for the thesis requirement for the Master of Science
degree in Industrial Engineering at the May 2007 graduation.

Thesis Committee:
Andrew Kusiak, Thesis Supervisor

Bradley Dixon

Paul Krokhmal

To my family and friends, especially Katie

ACKNOWLEDGMENTS

I would like to acknowledge the guidance of Prof. Andrew Kusiak. Without his leadership and supervision, the advancement of my knowledge in the area of data mining would not have happened. I would also like to acknowledge the guidance of Dr. Bradley Dixon. His expertise in the area of the treatment and care of kidney dialysis patients was instrumental in the synthesis of this document. I would like to acknowledge and thank Zhe Song, a fellow researcher at the Intelligent Systems Lab and IE PhD candidate for his help in formulating a research path. I would also like to acknowledge Ryan Donner, Ozgen Kilic, Ahmed Diken, and Robert Hamel for their help with feature selection and other data analysis.

ABSTRACT

Data mining is emerging as an important tool in many areas of research and industry. Companies and organizations are increasingly interested in applying data mining tools to increase the value added by their data collections systems. Nowhere is this potential more important than in the healthcare industry. As medical records systems become more standardized and commonplace, data quantity increases with much of it going unanalyzed.

Data mining can begin to leverage some of this data into tools that help clinicians organize data and make decisions. These modeling techniques are explored in the following text. Through the use of clustering and classification techniques, accurate models of a dialysis patient's current status are derived. The K-Means and Expectation Maximization clustering algorithms are utilized to generate homogeneous patient populations. Classification techniques, such as decision trees, neural networks, and the Naïve Bayes classifier are evaluated in terms of their accuracy performance. Time series aspects are also considered utilizing system identification techniques from control theory. Finally, cluster-derived classification models are tested for their cross-validation accuracy, as well as the generalizability to unseen testing sets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

A new and important paradigm shift is emerging in all fields of medicine. Research is becoming more focused on tailoring treatment to the individual patient. Identification of patient specific factors for successful treatment has focused on the genomic characteristics of the individual. It may be impractical to capture the genomes of each patient in a subgroup receiving a specific therapy, particularly ones with large populations such as hemodialysis patients. Of interest is whether these patient specific factors manifest themselves in more measurable characteristics. The complexity of treatment and therapy interaction may elude the abilities of traditional statistical analysis. A more robust method of modeling is necessary to test this hypothesis.

The potential benefits of individualized medicine are many. Drug dosages could be more tightly controlled on a patient by patient basis. Not only would this approach save money by reducing unnecessary dosage, but treatment could be improved as well. An approach based on data already collected would have certain unique benefits also. The cost of capturing an individual's genotype could be avoided if it could be demonstrated that advanced modeling techniques perform as well as treatment based on genotype information. Staff may be more inclined to ensure data quality if there are directly observable benefits to using data historians and medical records in this fashion. As decision support systems become more prevalent, clinicians would become more trusting in analysis and modeling technologies to help them do their jobs more efficiently.

As previously mentioned, new tools are necessary to capture more complex relationships between treatment, medications, and other patient-specific factors. Methodologies that require assumptions and controls may be confounded simply by the unpredictable nature of the human body. Data mining technologies build robust models that can capture these complexities using real data that is already being captured from a real process.

While individualized medicine is the overall goal, it is important to note the similarities between patients. By considering patients to be part of a homogeneous sub-population, the collective information from the sub-population can be used to enhance the treatment of the individual. Using a number of data mining methodologies, current grouping of patients can be improved. Chapter 1 deals with implementing techniques that automatically detect subgroups among patients – a data mining technique known as clustering. Two methodologies are contrasted in terms of their final output – the distribution of clustering parameters within the clustered populations.

Improvement of the state of the patient is of primary concern, and is the purpose behind treatment. In order to improve the state, one must first be able to accurately predict the state given past and current data. Theoretical response to a treatment may not hold true for much of a population. Therefore, data mining techniques are explored to test the accuracy with which they predict the current hemoglobin state in Chapter 2. Several techniques are presented, including decision trees, artificial neural networks (ANN), and the Naïve Bayes classifier. Feature selection with the genetic algorithm (GA) is also introduced as an important strategy to determine the optimal mix of attributes from a subset. Lastly, time-series data analysis techniques are reviewed, including temporal data mining and system identification techniques.

Finally, in Chapter 3, the implications of creating classification models based on clustered populations are analyzed. Using the clustered population derived in Chapter 1 and the feature selection and classification algorithms from Chapter 2, models are derived again to predict the current state of a patient's hemoglobin. A notion of the generalizability of the classification models is discovered when they are tested on a validation set and a randomly selected cluster population.

MOTIVATION

Data Mining

For medical purposes it is important to be able to interpret patient information and categorize similar people together. Mass production, economies of scale and standardization lead to the development of treatments, medicines, and processes that will have the best effects for numerous people. However, patients will still react differently to similar treatments. Personalized medicine seeks to match the right treatment to the right patient. Much research has focused on the effects of specific genes within a patient's DNA on their disposition to a treatment ([32],[33]). While this methodology has many merits, genetic data is unavailable for most patients. What is becoming more and more available is treatment and diagnosis data that can be utilized to extract hidden knowledge. This process is known as data mining.

Data mining was born from machine learning concepts in artificial intelligence. Figure 1 illustrates the iterative nature of deriving knowledge from a data set using this process. Data mining techniques have the ability to observe the complex nature in data processes. The key to utilizing the power of data mining techniques is to interpret and represent the derived knowledge in a meaningful way. Also key to any data mining project is the data preparation steps involved, which can influence the final outcome.

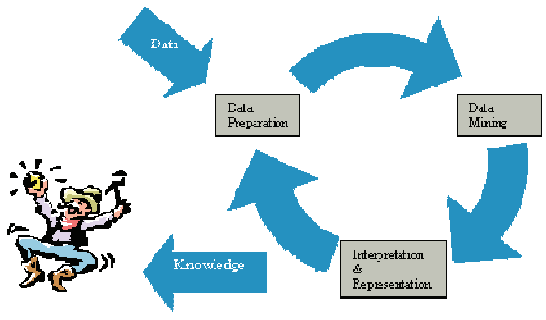

Figure 1: Mining for data mining gold

Data Mining Methodology

The basic approach follows the Knowledge Discovery in Databases (KDD) methodology outlined by Fayyad *et al*. [1]. Data mining is shown as only one step of the iterative KDD process:

1) Data selection: Choosing the proper subset of data to fit the data mining task.
2) Data preprocessing: Data cleansing, filling missing values, denoising.
3) Transformation: Further subset reduction, or change of representation.
4) Data mining: Applying a proper algorithm (based on the overall goal) to the data set.
5) Interpretation/Evaluation: Representing results in a useful way, assess validity of derived patterns.

An important step that coincides with data set selection is the formulation of a hypothesis. Can we predict the emergence of anemia in individual patients by using patient information stored in a database? Can we more tightly control the dosage of erythropoietin and iron, giving a proper dosage each dialysis and potentially reducing waste? This step is critical in order to understand the needs in forming a proper data set, as well as choosing representation and data mining algorithms.

Data feature selection is an initial reduction that typically occurs with the use of expert knowledge of someone within the domain. Domain experts have years of acquired knowledge that leads data miners in the right direction. For the purposes of this domain, expert knowledge of the dialysis process as well as important medications and complications (i.e., inflammations, comorbidities, and hospitalizations) helps narrow down the target data set.

Data preprocessing is extremely important and can effect the ultimate outcome of the data mining. Data sets must be in the proper format for the algorithm and software tools that will be used for the analysis. Missing data must be handled in some fashion. Some algorithms may require that continuous data be scaled between 0 and 1.

Transformation can involve some further steps in preprocessing. This can typically include the development of some derived features from the target data set, such as separating blood pressure into its systolic and diastolic portions. Ratios of inputs and outputs are common, as well as differences between measures. Another example of a common transformation will use some mathematical combination of features in order to determine a linearly separable decision boundary that may be non-linear in nature in a lower dimensional space. This technique forms the basis of the support vector machines (SVM) kernel function. SVM is a popular algorithm that requires a linearly separable decision boundary, which is not always inherent to many datasets.

Once the data set is prepared in the proper format for the chosen algorithm, the training of the model is undertaken. A sampling of the data set is typically used to form the model, and several methods of sampling are employed. A hold out method, typically 1/3 of the data, is used to differentiate a testing set.

Building classification models will yield classification accuracy - a percentage of properly classified instances from the test set. A confusion matrix shows the specific errors of the classification model. The model has classified four instances correctly as Low, and classified three instances as medium that should have been classified as low.

Table 1: Confusion matrix

Low	Medium	High	<=Classified as
4	3	0	Low
0	5	0	Medium
0	0	10	High

Figure 2 illustrates the specifics of the proposed methodology. From raw patient data the records are parsed into the static and dynamic groupings and perform the clustering algorithm to derive homogenous groupings. Further preprocessing must be

performed on the dynamic data to derive a standard representation for the data mining algorithm. This may involve further transformation of features, such as the condensing of erythropoietin or iron into a weekly dose. Once the dynamic data is prepared for mining, all patient records are gathered for a specific cluster and subjected to a rule-generating or data mining algorithm. These rules are then evaluated and tested for their generalizability to other data sets.

Testing for generalizability would involve the assignment of new patients to existing clusters and evaluating the patient with that clusters' rule base. Rule bases from one cluster may not apply well to patients of a different cluster, but in this thought, a new method of testing the validity of the clustering is derived. If rules for one cluster apply equally to the patients of another cluster, it may be concluded that the clustering was unnecessary and the clusters themselves may be superficial.

Cross-Validation

This testing technique makes maximum usage of the whole dataset. Each data record in this technique is used the same number of times for training and exactly once for testing. The dataset is typically partitioned into a certain number of "folds" of equal size. In the general case, *k-1* folds are used for training and the *k*th fold is used for testing. This process is repeated until all folds have been using for testing exactly once. As *k* increases, the process becomes increasingly computationally expensive. Therefore, for most datasets, a 10-fold cross validation is typically sufficient.

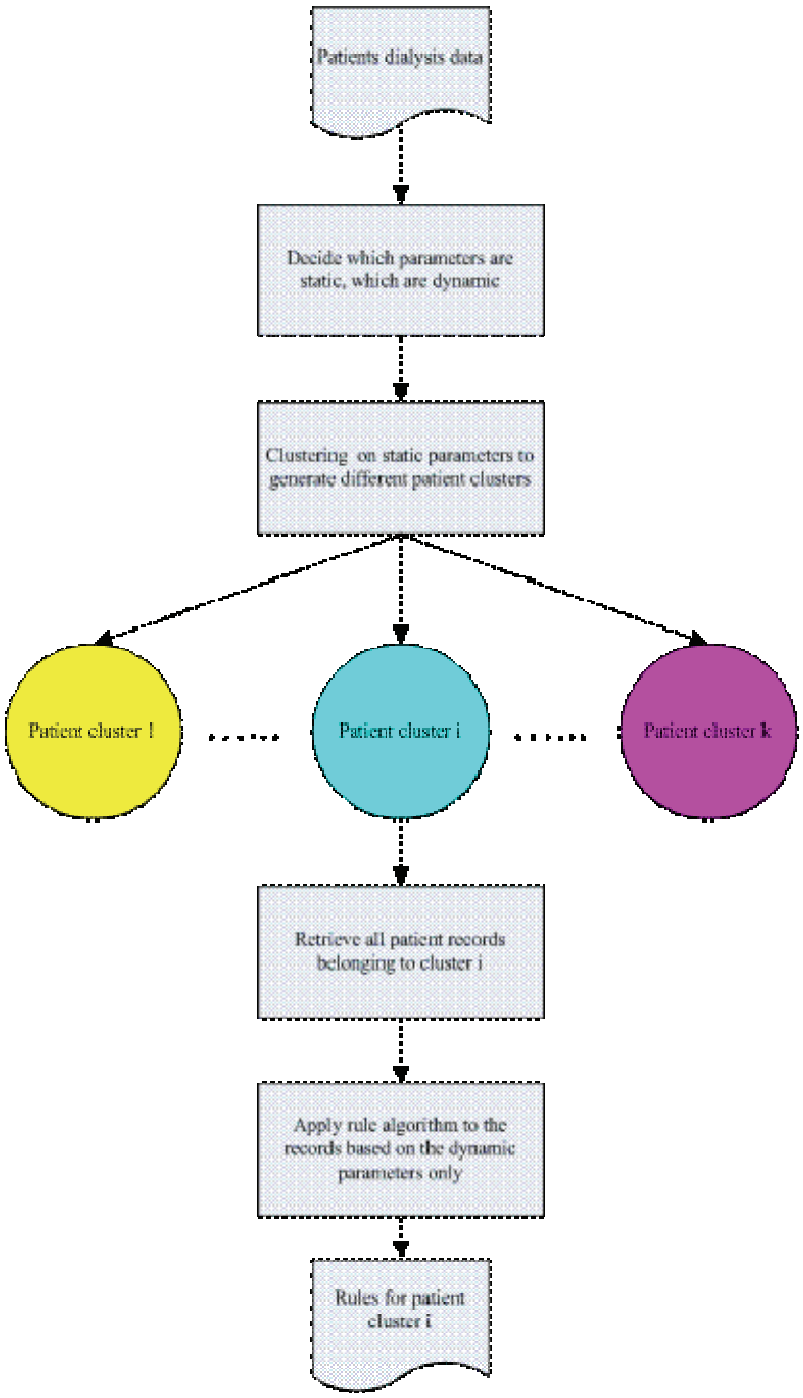

Figure 2: Flow chart of a proposed method for mining dialysis data

BACKGROUND

Kidney Function

The kidneys are part of the urinary system and are charged with removing waste from the bloodstream. They regulate chemical pH levels, as well as produce a number of important hormones that regulate other processes. Kidneys produce erythropoietin that allows the body to produce more red blood cells in conjunction with iron [30].

The process by which waste products and excess water leave your blood stream involves a complex chemical exchange. Once these products are filtered from the blood, they enter the urinary system and are stored in the bladder until they are removed through urination. A healthy human with both functioning kidneys is said to have 100% of their kidney function. As age or disease sets in, humans can function normally with more than 25% of their kidney function. As kidney function approaches 10% to 15%, renal replacement therapy is required. Renal replacement therapy is typically categorized into transplantation and dialysis [30].

End Stage Renal Disease

Kidney failure is defined as the short-term or long-term damage of the kidneys. This condition leads to the loss of normal functioning of the kidneys. Failure can be abrupt in onset, or chronically degrade over time. Acute renal failure can be caused by heart attacks, kidney damage, decreased blood flow, and complications from certain medications. Chronic renal failure is brought on by diabetes, chronic high blood pressure (hypertension), lupus, and kidney disease. When the kidneys finally fail completely, the condition is known as end stage renal disease (ERSD). Many symptoms accompany this final condition, and it is typically diagnosed through blood tests, urine tests, and many other examinations. Treatment is typically rigorous, including diet monitoring, hospitalizations, and most notably transplantation or dialysis. Candidates for transplantation must meet certain criterion in order to undergo the operation [28].

Kidney Dialysis

There are many variables that make up a dialysis treatment. There are several types of dialysis apparatus utilized to filter the blood, along with different types of accesses. Dialysis essentially mimics the function of the healthy kidney, and medications are substituted in the cases where the kidneys have also stopped producing certain hormones.

Anemia is defined as a deficiency of red blood cells or hemoglobin, which translates to a decreased ability of the body to transfer oxygen to the cells. Anemia can occur in one of three ways: excessive blood loss, excessive cell destruction, or deficient cell production [27].

Iron supplementation is common for dialysis patients as blood loss generally occurs during the course of treatment. Oral iron is not as easily absorbed as iron delivered intravenously, though complications from some forms of intravenous (IV) iron have been discovered in a small percentage of patients (<0.1%) [26]. Iron is required by new red blood cells, along with erythropoietin in a process known as erythropoiesis. Red blood cell production occurs mostly in the bone marrow, specifically in the leg bones until about the age of 25, and life-long production occurs in the vertebrae, sternum, and ribs [31]. Erythropoietin is produced in the kidneys, but production can be slowed or stopped completely in patients with kidney disease. A synthetic replacement is available in the form of Epoetin, and is administered to patients suffering from a variety of diseases that may cause anemia

Best Treatment Practices

Controversy abounds in relation to the development of guidelines for controlling hemoglobin levels within certain ranges. For most of the 1990s, guidelines stated that the target hemoglobin range for patients should be between 11 and 12 g/dl for all patients with chronic kidney disease. This was range was increased to 11 to 13 g/dl following the

2006 Kidney and Dialysis Outcomes Quality Initiative (KDOQI) anemia guidelines release. Mean weekly usage of erythropoietin has been on the rise and is expected to continue to do so with an accompanying increase in hemoglobin levels. However, this increase has led to no significant effect on the numbers of hemodialysis deaths.

Drüeke *et al.* [5] found that while general health and physical function are improved, normalizing hemoglobin levels (13.0 to 15.0 g/dl) does not decrease the risks of cardiovascular disease in patients. The population included patients from 94 dialysis centers in 22 countries. Clinicians were instructed to follow current clinical practices. The patients were split into two groups, with the second group receiving only partial correction (10.5 to 11.5 g/dl hemoglobin). Complete correction did not decrease the likelihood of a first cardiovascular event and more patients in group 1 required dialysis. Hypertension and headaches occurred more frequently in group 1.

Treatment of anemia is certainly complex, and a prime candidate for the beneficial analysis that data mining has to offer. Capturing the interaction between medication, dialysis treatment, and patient condition complicates many studies attempting to isolate independent variables. This contributes to the inconclusive nature of the results of many studies, and the subsequent difficultly in determining proper treatment practices.

According to the National Kidney Foundation's KDOQI anemia guidelines, target hemoglobin/hematocrit should fall in the low normal range. This happens to be 11 g/dl (33%) and 12 g/dl (36%) hemoglobin (hematocrit). This guideline is based on evidence, but does not take into account the differences in physiology of men and women. Survival is found to be lower in patients that have a lower hematocrit level (30% to 33%). According to Besarab *et al*. [7], mortality decreased in patients that maintained normal hematocrit as compared to a control group that did not attain and maintain a normal hematocrit level. According to Xia *et al*. [8], hospitalizations were fewer in patients with hematocrit levels between the target range.

CLUSTER ANALYSIS

One of the common tasks in patient care is to identify what differences exist between patients that cause varying results. Categorizing of patients already occurs, but may focus only on a limited number of factors to determine proper treatment, or may focus too heavily on population-based statistics. An alternative approach would be to utilize data mining techniques in order to properly group patients together based on similarity metrics over a wide variety of historical treatment data. Categorization can be accomplished using many types of data including demographics, test results, diagnoses, etc. Consider the following treatment model:

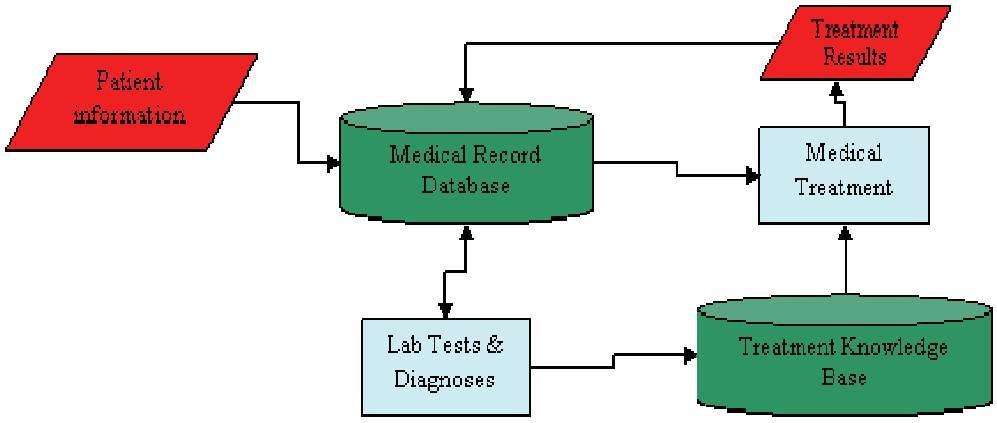

Figure 3: Treatment Information Flow diagram

A patient receiving dialysis has accumulated a data record consisting of many items including medications, previous treatment history, demographics, etc. These factors and countless others all have an influence on the success of a medical treatment, and also implies that patients will react differently to similar treatments.

Data mining analysis can be very sensitive to noise and erroneous data. Rules and patterns derived from large, heterogeneous data sets may be plentiful, but the quality of these rules may be suspect. Human data is a prime example. Each person in a database can be viewed as their own system – taking in a set of inputs (such as medications, diet, sleep, etc) and producing unique outputs such as vital signs and test measurements. Data mining using only single patient data may not contain enough information to draw generalizable outcomes. On the other hand, taking information from a large number of patients may overwhelm important patterns.

Utilizing an entire population of kidney dialysis patients for analysis may not address these differences. Often when an excessive amount of data is utilized, excessive meaningless patterns are inferred [2]. Through clustering we can understand how each patient relates to one another, assigning them to groups of similar patients. Clustering is a data mining technique that automatically assigns instances (in this application - patients) of a data set to groups of similar instances. This assignment can be made according to any number of parameters including categorical and numeric data. Static parameters such as demographics can be utilized to derive these groupings. Demographics are often analyzed to determine if ethnicity, race, or gender predispose patients to a particular disease.

Related Literature

Albayrak and Amasyali [15] implemented a clustering method to assign patients to different clusters of thyroid disease. The researchers contrasted the cluster assignments made by a fuzzy c-means clustering with that of a hard k-means clustering. The clustering served as an unsupervised classification method, labeling the patients according to their particular characteristics.

Bensmail and Meulman [19] proposed a methodology very similar to the EM algorithm that utilized Bayesian theory to construct clustering-based classifiers in a

number of domains, including categorization of diabetes patients. The mixture-models were evaluated using a Bayes factor to simultaneously evaluate multiple models.

Clustering is a commonly used term in medical studies [20]. It should be pointed out that though the concepts described here are inherently similar, there is a difference between clustering based on statistical distribution of patients in a certain population and automatic cluster detection. Though some cluster detection methods make use of the underlying distributions (EM algorithm, which is discussed in depth later), cluster assignments are not dependent on a single parameter alone.

Clustering Analysis – K-Means Algorithm

One of the most common clustering implementations involves the K-Means Algorithm. This algorithm follows the following steps:

Equation 1: K-Means Algorithm

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

This random initialization is rather important, as the clusters will not always converge to similar positions over a number of trials. Choosing the number of clusters can also be an issue. Successive trials can be run to keep track of the sum of squared

errors between centroids. The value of k that minimizes this value can be selected as the proper number of initial centroids.

Two-feature examples of k-means algorithm implementations are useful to understand the movement and path of the centroids from one step to the next. Being easy to visualize, the reevaluation steps are illustrated as follows:

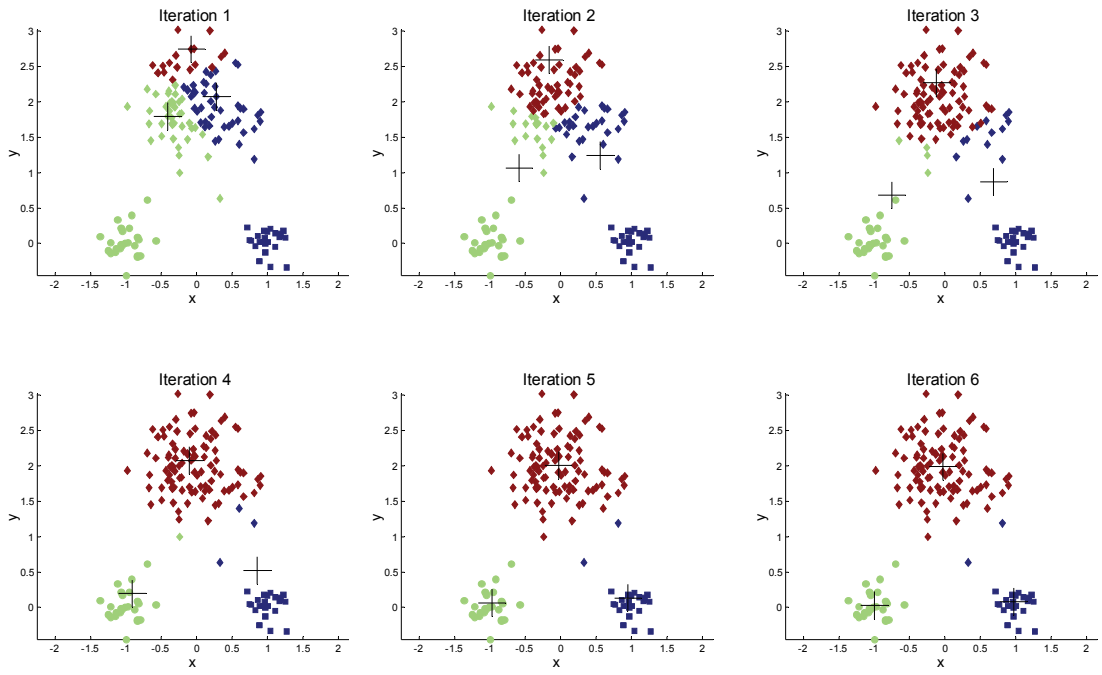

Figure 4: Iterative steps of K-Means Clustering

Figure 4 illustrates the simplicity of the algorithm, as the centroids converge to their final values in Iteration 6. This may not always be the case

In higher dimensional space, the clustering becomes more difficult to visualize. The key to accomplishing the clustering task is to choose an appropriate distance metric. The proper metric depends on the type of data in the dataset. Euclidean distance is a typical distance metric when dealing with numeric variables on defined ranges.

Hamming distance is a good measure when dealing with Boolean data as it simply counts the number of bits that are dissimilar.

Initial Clustering Analysis – Feature Selection

The overall goal of the clustering analysis is to define groups of similar patients. How these grouping ultimately affect the classification outcome must also be considered. To this end, a feature could be derived from the data that defines the patients in terms of how well they respond to EPO therapy. According to the literature and clinician experience, race and ethnicity play an important role in determining a patient's genetic predisposition to survival and certain treatments [12]. Certainly the response between male and female will be different and must be accounted for. It has been demonstrated that females are prone to developing higher erythropoietin resistance than men [13]. Patients will fall along some sort of continuum of response. The patient's age and number of years on dialysis will also be considered important. Younger patients may be overall healthier than older patients, or patients whom have been on dialysis for a number of years.

Table 2: Clustering features

Parameter	Description
ClinicIDNumber	ID number of treatment clinic
Ethnicity	Ethnic decent of patient (Hispanic, Non-Hispanic)
Race	Racial group (i.e. White, Asian, Indian)
Gender	Male or Female
CurrentAge	Age at time of analysis
YearsOnDialysis	Number of years on dialysis

Some important questions to answer by performing this cluster analysis are:

1) How important will the clustering consider gender, ethnicity, and race when determining the similarity between patients?

2) How will age and years on dialysis interact with the other features, specifically gender and race?
3) How will different centroid initializations affect the resulting clusters and what strategy can be employed to avoid these affects?

Clustering Results

As k is increased, a trade-off occurs between sum of squared error and significance of the size of the clusters. The final size of the total dataset after all patients were eliminated from consideration was 1,057 patients (596 male, 461 female).

Table 3: Sum of squared error between clusters for different values of k

k	SSE	Largest Cluster
2	3441.81	55%
3	3337.95	38%
4	2699.44	37%
5	2483.87	35%
6	2472.91	35%
7	2296.03	35%
8	2237.51	35%
9	2138.63	35%
10	2019.16	24%
12	1674.79	27%
14	1610.42	20%
16	1398.22	18%
18	1378.82	18%
20	1342.1	18%
22	1315.38	19%
24	1218.88	17%
26	1145.05	17%
28	1139.72	17%
30	1064.29	13%
40	760.03	8%
50	707.07	6%

Significant cluster size is desirable – one should contain at least 10% of the total population (100 patients). Using this size metric, the derived patient clusters will have a

sufficiently large population to build data mining models. Table 3 shows the trade-off between SSE and largest cluster size as k is increased. When $k = 10$, we reach a sufficient point of significance for SSE and cluster size. The clusters for $k = 10$ show some very interesting characteristics. For this particular clustering of patients, Table 4 shows the purity of the clusters in terms of the percentage of males. For the most part, the clusterings are fairly pure in terms of gender, with the exception being cluster 8. Cluster 8 is a small cluster, containing only 20 patients of mostly Mexican descent. These patients are of a diverse age range, and vary greatly in terms of how long they have been on dialysis. It would appear that the k-means algorithm chose to group these patients primarily on their ethnicity.

Table 4: k = 10 Cluster Purity in terms of Gender

Cluster	% Male
Cluster0	0.086666667
Cluster1	0.960227273
Cluster2	0
Cluster3	0
Cluster4	1
Cluster5	0
Cluster6	1
Cluster7	1
Cluster8	0.45
Cluster9	1

For most of the clusters, age and the number of years on dialysis are not important parameters in terms of highly defining the overall cluster. Figure 5 shows a typical age distribution, it is fairly representative of the clusters overall, with the exception of Cluster 7. Figure 6 shows the distribution of cluster 7 patient ages. The mean for this sub-population is approximately 44 years (mean of overall population is 66 years). From the information in Table 4, cluster 7 is

pure in terms of gender with 100% males. The vast majority of patients comes from clinic 1 and has been in treatment for less than 2 years. This cluster is also quite diverse, with seven different ethnicities represented.

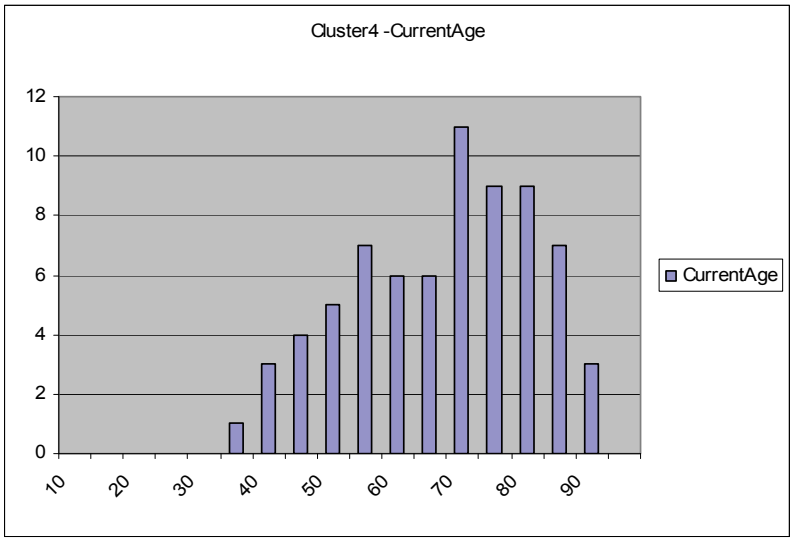

Figure 5: Typical Age distribution for most k=10 Clusters

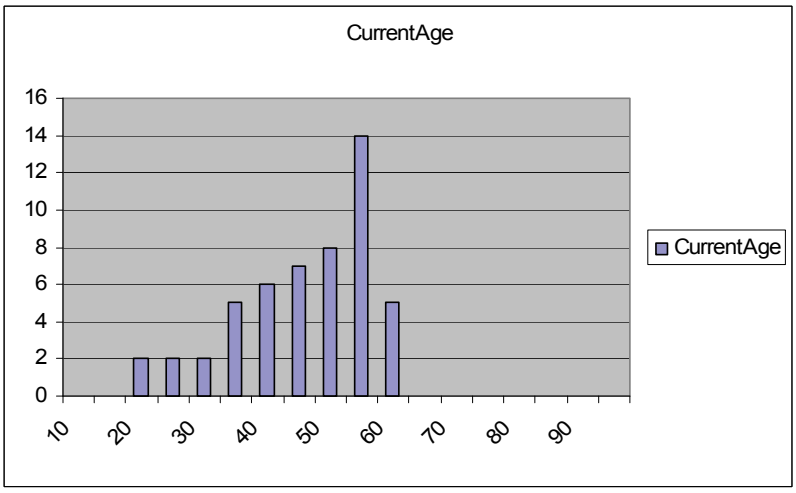

Figure 6: Cluster 7 age distribution

Clearly there is potential for interesting groupings that may have a significant affect on the nature of classification models derived from these particular subgroups. Figure B.1 though Figure B.40 illustrate the remaining distributions for the k-Means clusters.

Initialization of the K-Means algorithm can have a significant effect on the resulting clusters. It can be just as important to properly initialize the clusters as choosing the number of clusters. Several strategies can be used to overcome this disadvantage. Studying the cluster assignments of several trials of the algorithm using varying random seeds is a viable and often implemented strategy. Another alternative is using an algorithm such as the EM algorithm that is not subject to this disadvantage.

Expectation Maximization Algorithm

Instead of choosing k in such a subjective manner, the expectation maximization (EM) algorithm can be used to automatically choose the number of clusters based on probability distribution estimation. The EM algorithm form what are known as mixture models – models that describe the data using statistical distributions. The steps of the algorithm are as follows:

1. Select an initial set of model parameters (randomly or otherwise)
2. repeat:
3. Expectation Step: For each object, calculate the probability that each object belongs to each distribution, i.e., calculate *prob(distribution j*$|x_i,\Theta)$.
4. Maximization Step: Given the probabilities from the expectation step, find the new estimates of the parameters that maximize the expected likelihood.
5. until: The parameters do not change (or change below some threshold)

In practical terms, the distributions for the features in the dialysis patient set are unknown, thus we cannot directly calculate the probability of each data point. Step 1 of the EM algorithm overcomes this fact by estimating the parameters of the distributions.

Consider a set of points generated from a Normal distribution. The probability of obtaining this particular set of points is the product of their individual probabilities. Using the probability density function for a Normal Distribution, this function would be defined as:

$$prob(\chi, \Theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Given a data set, we could instead estimate the model parameters using a likelihood function:

$$likelihood(\Theta, \chi) = L(\Theta, \chi) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

This likelihood function is typically transformed using the log function in order to work with numbers on a much higher order to magnitude, since the probability will likely be very small. The solution of this equation is the parameters μ and σ that maximize the likelihood function.

EM Algorithm Results

The EM clustering algorithm chose to partition the data set into six clusters of varying size and composition. Most clusters contained over 100 patients except for cluster 2, which only consisted of five very particular patients.

The resulting clusters are much less pure in terms of gender than the clusters derived from the k-Means algorithm. Interestingly, the clusters were mostly comprised of approximately 50% men and women. Men and women are known to react quite differently to numerous drugs and treatments, but these clusters are unaware of the ultimate goal of finding a pure sample population with which to derive highly accurate classification models.

Table 5: EM Algorithm gender cluster purity

ClusterID	% Male
0	0.660714
1	0.503425
2	0.6
3	0.598253
4	0.396552
5	0.590909

An interesting result of the EM clustering was an emphasis on the continuous attributes – age and number of years on dialysis. The clusters appeared to emphasize some of the differences between groupings on these two particular parameters. Cluster 2, as mentioned previously, was quite an anomaly. The following figures show the distributions of the continuous variables for this subgroup.

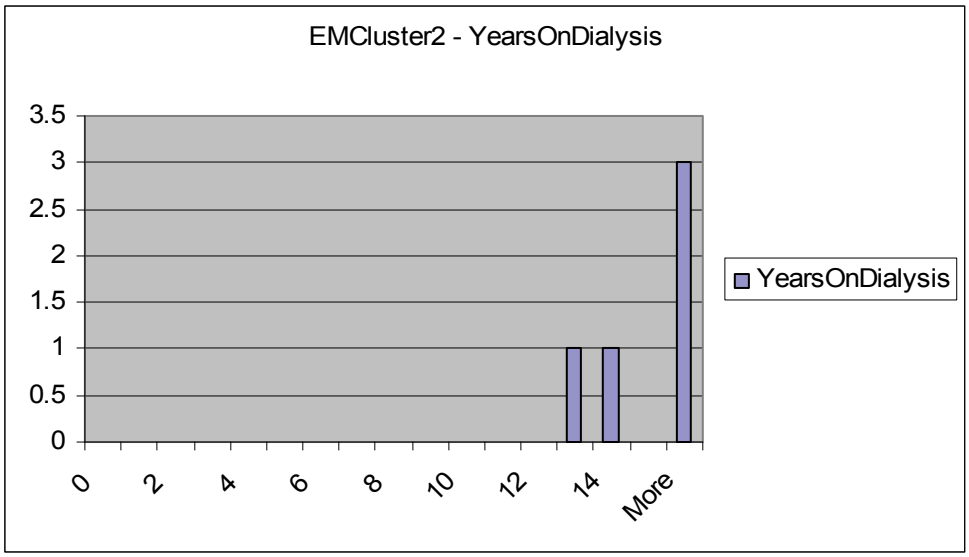

Figure 7: Number of years on dialysis for EM Cluster 2

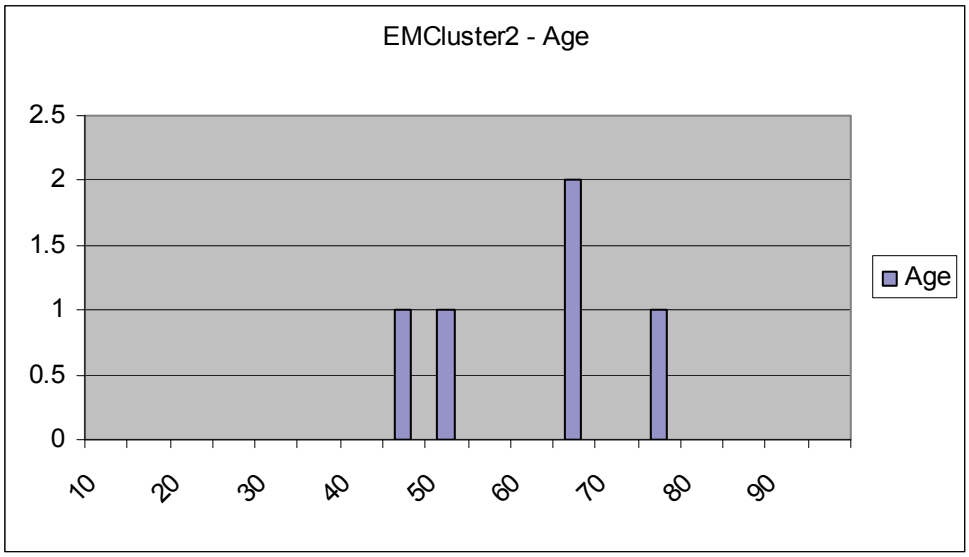

Figure 8: Age of patients for EM Cluster 2

These five patients have been receiving dialysis longer than any other patient in the dataset. Median patient do not survive on dialysis much longer than two to three years (see Figure A.1). It is a quite interesting result and may show the inherent bias in using the EM algorithm when confronted with both categorical and numeric attributes. This is intuitive, as the Weka implementation of the algorithm converts categorical values to separate binary attributes. This type of bias may be of clinical importance to recognize the significance of continuous attributes. Figure C.1 through Figure C.15 (Em Clustering Distributions) illustrate the remaining age distributions for the EM clusters.

Implications of Adding Additional Features

Of primary importance in deriving an efficient and meaningful clustering is incorporating some metric that is an indication of the patient's reactivity to the drug therapy. More specifically, it is of interest to understand each patient's response in hemoglobin to a change in EPO and its equivalents. This goal can be achieved in numerous ways. Simple metrics would use as a clustering feature the degree of

variability in the patient's dosage of EPO over the course of treatment. This is only practical when EPO and its equivalents have already been delivered, and cannot be evaluated on patients that do not have a treatment record with the particular drug.

This is an issue in data mining in general – how to build useful models when there is a lack of historical data. Generally there is not clear answer for this problem, but in a clinical setting, inferences can be drawn from the patterns of other patients. This is done on a very regular basis in current practice.

In general, there are few objective methodologies to evaluate the "goodness" of one clustering over another. Purity and other measures are only applicable if those attributes are desirable. The potential usefulness of the clusters, in this particular application, will be explored further in Chapter 3. Future studies may make better use of techniques to optimize the feature set used in clustering, but some benchmark of cluster usefulness must first be established.

Clinical Applicability

This application of clustering algorithms to efficiently and automatically group patients can have far reaching implications. Feature selection and data type was shown to be important in conjunction with the algorithm applied. Patient populations such as hemodialysis patients are receiving treatment for a very diverse set of reasons. These algorithms can be applied to more accurately group them together for common "treatment pools". This combined with modeling of dynamic patient outcomes can lead to better treatment overall. K-Means algorithm is an extremely efficient choice for performing rapid clustering applications with a variety of datasets. Though this knowledge would still not move treatment completely towards individualization, there would be a strong potential to increase treatment efficiency simply by considering a more complex notion of similarity. As the number of treatment pools discovered amongst the total population increases, the number of patients in each pool would decrease.

MODELING HEMOGLOBIN RESPONSE

The dialysis data set can be separated into static and dynamic portions. Demographics and other patient information such as diagnoses can be considered static, while test results, dialysis run information, and vitals are considered dynamic. Using the static parameters to derive groupings, we can then generalize a treatment pattern for these patients based on their changing responses. The number and size of clusters can be manipulated to an optimal number, and results from different modeling can be compared using prediction accuracy.

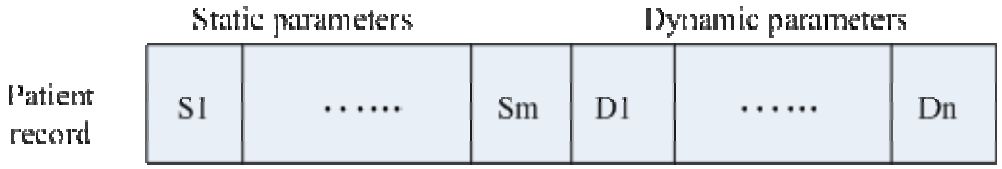

Figure 9: Patient record can be divided into different parameter types

Many tools are available to derive the expert rules using data mining. The knowledge discovery field has borrowed many tools not only from artificial intelligence, but also information theory and statistics. The decision tree algorithm is a common method used to derive rules. The final output is also rather intuitive and easy to interpret. A common implementation of the decision tree algorithm uses a concept from information theory known as "information gain" to evaluate each parameter in a data set in terms of its importance in predicting the final outcome. The following image is an example tree developed from a small analysis of a single patient's kidney dialysis treatment data. The tree was derived using the Weka data mining environment.

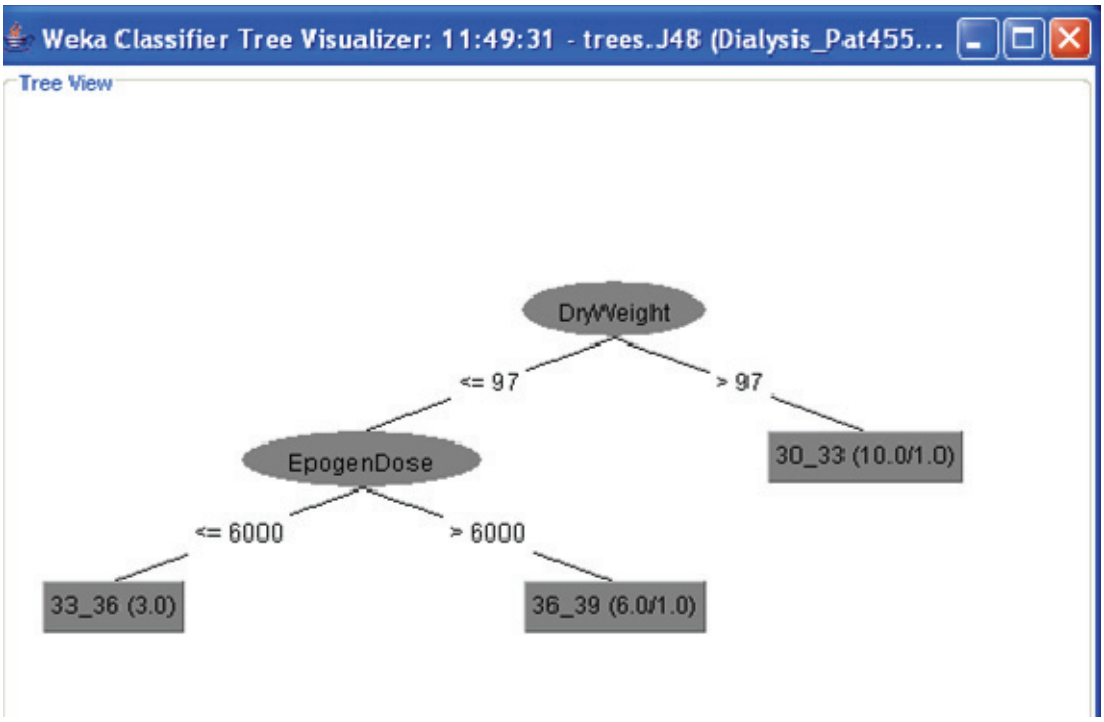

Figure 10: Weka Decision Tree

This model for this patient was built to predict the outcome of the hematocrit test for this particular patient. Hematocrit is the ratio of the volume of red blood cells to the total blood volume. Dialysis run data was incorporated along with the dosages of Epogen and iron to predict a hematocrit test result that occurred approximately every two weeks. The test results were discretized to intervals of three starting above 30 and below 39. Each box is known as a leaf node in the decision tree. The circles are parameters of the dataset

The results of this analysis may be spurious at best, but it can be used as an example of the potential output. Indeed, the formulation of such results may be one of the other challenges to the application of data mining in many other domains. The hope

is that implementation of the clustering analysis preprocessing step may lead to better rules, and this can be verified in contrast to rules generated without clustering.

Clinical trials seek to establish a relationship between a given treatment and a specific response, such as the dosage of erythropoietin and the level of hemoglobin. Guidelines are set out by the National Kidney Foundation on how best to treat the emergence of anemia in patients. It is of definite importance to contrast the generated rules with the set guidelines used by clinicians. Most of the generated rules, however, will be more complex and may include parameters that may not make intuitive sense in clinical terms.

Decision Tree Classification

Induction using decision trees is particularly popular for domains that require an easily expressible output of rule sets. Decision tree algorithms are supervised learning techniques (i.e. the outcome is known while training) that typically use binary splits of data set features. Figure 4 illustrates a decision tree that would be used to decide to play tennis or not based on the temperature and the outlook.

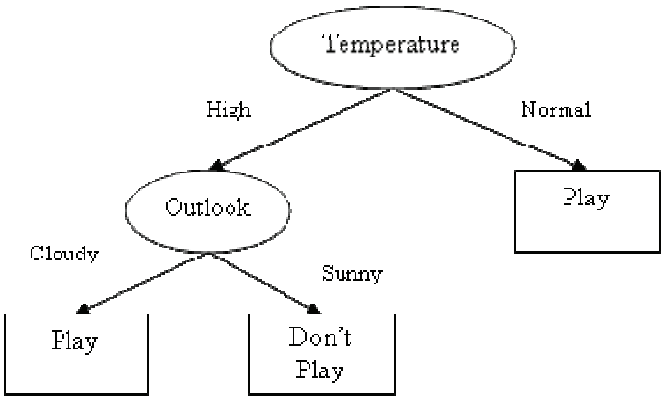

Figure 11: Decision tree for playing tennis

Implementations of the decision tree algorithms use many methods to determine the optimal feature to split and the number of splits to make. The number of splits may depend on the variable type, such as nominal attributes. Continuous attributes can use a binary comparison according to some split value that is determined in a similar fashion to automatically determining the optimal splitting feature. The optimal split value can be evaluated in a similar manner to the attributes. A metric for doing so is presented next.

Several methods for evaluating feature impurity exist and are implemented in decision tree algorithms. The metric that typically produces the best results is entropy.

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i \mid t) \log_2 \; p(i \mid t)$$

$Where\ c = \text{Total \# of classes}$

$t = \text{Current decision node}$

$i = \text{Class}, i \in C$

Entropy measures the randomness of a feature in comparison to a given class. The measure is akin to the entropy measures of thermodynamics, but is mainly used as an implementation from information theory, a branch of computer science [2]. The impurity, in this case entropy, is then compared to the impurity of the parent node. The feature that has the largest difference from the parent node is greedily chosen as the best splitting feature.

Dataset Description

A comprehensive dataset with over 3,000 patients was used for this research. This dataset included dialysis run information, HIPAA compatible demographic information, medications, and hospital visits. Dialysis treatment was performed in a number of clinics across the country, but the clinics are identified only by an ID number in the dataset. Data quality issues reduced the usable number of patients over a number

of iterations. The number of patients is still quite large in comparison to most clinical trials.

Data quality is a large issue in data sets of this nature. Standardization of fields may be minimal, and the different clinics that enter data into the common database may place a higher value on different data. Many fields are sparsely populated, making it difficult to estimate missing values. Missing values must be handled in some manner for most algorithms.

Patient Population

For the total patient database population, average age for the dialysis patients is 66 years with a standard deviation of 15.6 years. Figure 5 illustrates the distribution of the patient age. The median for this particular population is 68 years.

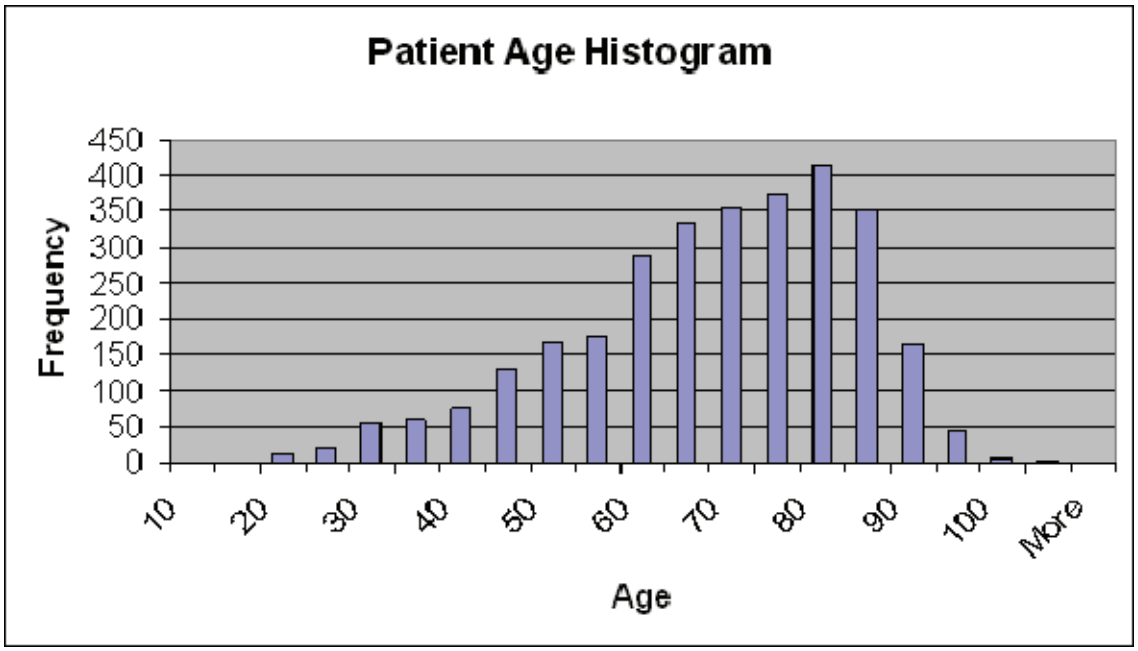

Figure 12: Age distribution of total dialysis database population

Mortality is high in patients with end stage renal disease; many patients do not survive more than a few years on dialysis. An important feature of this dataset is the date of first dialysis. Dialysis run data is not comprehensive from this initial date, so an assumption will be made that dialysis is performed consistently through the years.

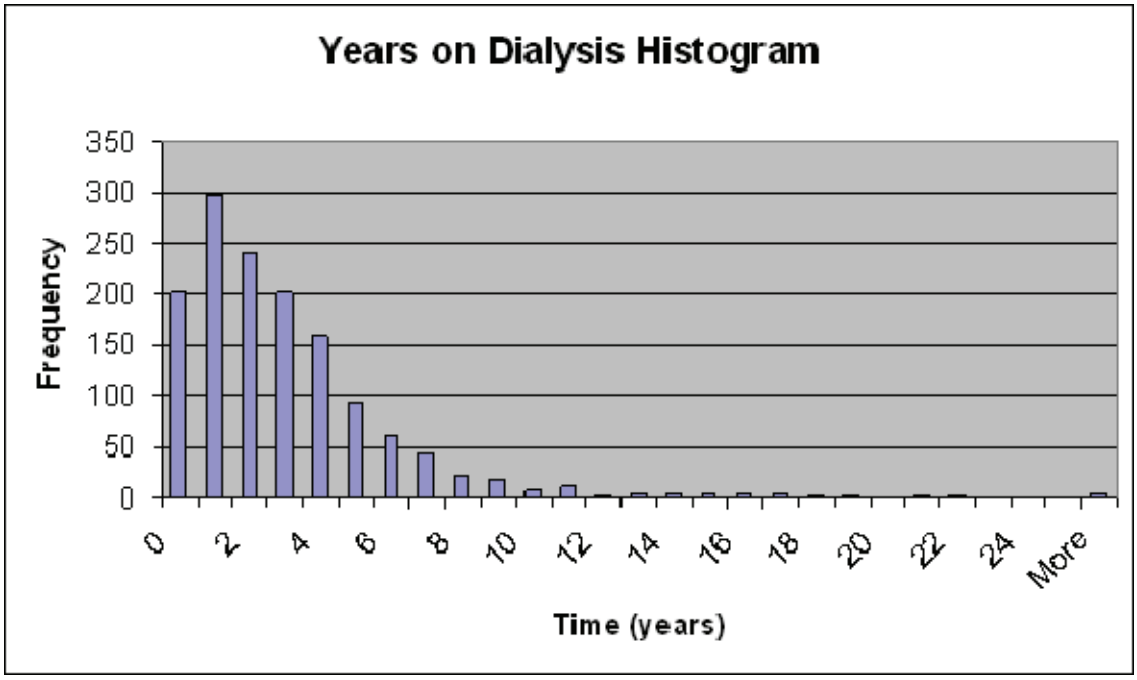

Figure 13: Distribution of the number of years on dialysis

The date of first dialysis figure was not available for all patients. Figure 6 is based on nearly 1,400 patients. 90% of these patients have received dialysis treatment for less than seven years (see Figure A.1). Mortality data is not available for any patients in the database. This subgroup of patients is described by a mean of 3.00 with a standard deviation of 3.14 and a median of 2 years.

Attribute Selection using Expert Knowledge

Insight from a domain expert is essential, as the expertise in the treatment of dialysis patients is not trivial to acquire. Candidate variables were selected on the basis of their influence on some biological factor. These biological factors were determined through discussions with Nephrologist Dr. Bradley Dixon and through review of the hemodialysis literature. In some cases, these variables would need to be transformed, as was the case with Epogen dosage (See Derived Variables). Many important factors influence the uptake of erythropoietin, including gender, ethnicity, inflammation, comorbidities, etc. Gender, in particular, appears to modify the receptiveness to EPO. (Ifudu, 2001) One particular benefit of working with such a large data set is the freedom to also include variables that may not be considered important factors in the classical clinical setting.

Derived Variables

Epogen dosage changes in response to lab values of hemoglobin and hematocrit. However, the effectiveness of Epogen dosage and other medications may be influenced by the fluctuation of the patient's weight. An important concept is the patient's dry weight.

Definition 1: Dry weight is the "ideal weight" of the patient when excess fluid has been removed

This target weight can fluctuate based on weight gain or loss of the patient, but it gives the clinician a target to shoot for. Depending on the effectiveness of the dialysis treatment, there will be a degree of variability of the actual weight difference between pre- and post-dialysis. The post-dialysis weight is then used to derive the ratio of dosage in terms of the patient's weight:

Definition 2: The Epogen/weight ratio is defined as the quotient of the weekly Epogen dosage divided by post-dialysis weight. (ex. 50 units/kilogram)

Practical reasons existed as well for choosing post-dialysis weight as opposed to dry weight. Post-dialysis weight was collected on a much more consistent basis. Much more variability was found in post-dialysis weight as well. Dry weight is more prevalently used as a metric for determining the dosage of Epogen, but it would be valuable to study the effect of actual weight removed.

Time on dialysis is simply duration of time the patient was treated with dialysis. This attribute was not measured directly, but estimation was made based on the starting and ending times recorded in the database. This process of recording these times may not have been automated in some cases, and therefore there may be a degree of error in the time on dialysis measures.

Lab Result Values

An important preprocessing step was undertaken to rectify the periodicity of certain measures in the dataset. Dialysis runs are performed three or four times per week with a mostly constant Epogen dosage. Lab values, such as the results of blood tests that check for the concentration of hemoglobin and hematocrit, are taken at different intervals, sometimes once or twice a month. These values are utilized for the following dialysis runs, so for the purposes of the dataset they will be held constant until a new lab value is found in the database.

Decision Variable

The techniques utilized for building our models were categorized as "supervised" learning methods. Ultimate outcomes of test sets are known; therefore classification accuracy can be measured. This is the distinction between supervised and unsupervised

learning. The most prevalent measure of anemia in a patient is the laboratory hemoglobin test; therefore this was chosen as our decision variable. As previously mentioned, National Kidney Foundation guidelines have set a target value of between 11 g/dl and 12 g/dl, regardless of sex. Values that lie in this interval are hereafter considered in the "normal" range. Values below 11 are to be considered "Low", and values above 12 are to be considered "High". This discretization is necessary for most classification algorithms. Alternative discretizations could choose any size granularity over the continuous range of values. However; as the number of bins increases, the population in each bin decreases and training and testing accuracy may decrease. This will be taken into account when deciding what level of discretization will be used, and what the clinical ramifications will be.

Genetic Algorithm Feature Selection

A test set is derived based on the information from the dialysis runs of twenty-two patients. This set includes approximately 3,600 instances (runs) over seventeen features and the hemoglobin outcome. These patients were selected at random from the entire population, and do not contain any overriding similar characteristics.

Table 6 shows the input variables initially selected. "AccessType" refers to the vascular dialysis access employed in a particular patient. This variable is not static: a new access can be placed and used for treatment while an older access is phased out. "TimeOnDialysis" is the length of time in minutes of a particular dialysis run. "InfedDose" is the dose of intravenous iron given to the patient. Like EPO, it is associated with a unique dialysis run. "FerritinLabResult" is an indicator for the amount of iron in the blood stream, an important predictor of anemia. "MCVLabResult" is the results of the mean corpuscular volume test, which indicates the average volume of single red blood cells, also an indicator of anemic condition [21]. "AlbuminLabResult" is the result of the lab test for albumin, an indicator of inflammation.

Table 6: Genetic Search Initial Features

AccessType	{AVfistula, GortexGraft, PermCatheter, TessioCatheter, VectraGraft}
TimeOnDialysis	Continuous Number
InfedDose	Continuous Number
FerritinLabResult	Continuous Number
MCVLabResult	Continuous Number
AlbuminLabResult	Continuous Number
WBCLabResult	Continuous Number
StartSittingBP_systolic	Continuous Number
StartSittingBP_diastolic	Continuous Number
StartSittingBP_difference	Continuous Number
EndSittingBP_systolic	Continuous Number
EndSittingBP_diastolic	Continuous Number
EndSittingBP_difference	Continuous Number
EPO/KG	Continuous Number
Gender	{Male, Female}
Ethnicity	{1, 16, Non-Hispanic}
Race	{1, 16, 3, 4, 5, White}

"WBCLabResult" is the results of a patients white blood cell count. In terms of a biological indicator of anemia, an increased white blood cell count is an indicator of an inflammatory condition as well [16]. Also included as attributes are several indicators of the patient's blood pressure, both before and after the procedure has taken place. "EPO/KG" is the dosage of Epogen divided by the patient's post-dialysis weight. "Gender" is the patient's sex – male or female.

Ethnicity in this data set typically refers to whether a patient is from Hispanic or non-Hispanic background, while "Race" is defined as a particular subgroup within an ethnicity group. Some of these categories are defined by numbers, but the mapping of these attributes to their real world counter-parts is unavailable at this time. These patients were still considered in the data set in order to evaluate a larger population for both clustering and classification tasks.

These attributes were chosen based on the clinical knowledge of a staff nephrologist at the Veteran's Administration Hospital in Iowa City. This

particular selection is not all-inclusive: there are many attributes that were not selected that could be chosen as viable candidates as well. Luckily, data mining offers tools to evaluate candidate variables under some specific frameworks.

A wrapper is defined as a data mining attribute selection tool that considers the classification tool in making its selection. Typically defined as a heuristic search, a wrapper feature selection tool can use as its evaluating criteria the classification accuracy of a particular applied algorithm in evaluating combinations of features. One particularly powerful tool incorporates the Genetic Algorithm to derive an optimal combination of feature subsets.

Genetic Algorithms (GA) mimic the mechanisms of gene combination over generations of evolution. Parents (an initial population of individuals) combine their genes to produce children in a new generation. The children's ability to pass on their genes to subsequent generations is termed "fitness". The fitness of an individual is typically some relevant heuristic – in the case of the wrapper feature selection tool, the fitness function is the classification accuracy of a decision tree derived from a particular subset of features. Consider the following example.

A dataset with ten input features is selected for evaluation with a GA wrapper using the C4.5 decision tree algorithm for the fitness function. An initial population is generated at random with the following format:

Individual 1 : 1001011011
Individual 2 : 0101110110
Individual 3 : 1000100111
Individual 4 : 1011000011

Each bit represents a one of the 10 features from the input data set. A bit is set to one if that feature is to be included in the subset, zero is not. At this point features are randomly chosen to be combined:

Individual 1 : 1001011011
Individual 3 : 1000100111
- - - - - - - - - - - - - - - - - - -
Individual 2 : 0101110110
Individual 4 : 1011000011

For each pair, a cross-over point is also randomly chosen. The relationships will produce two offspring each. There is also a small chance of mutation – selecting a bit to change value. Child 42 contains one mutation. At this point each of these children will have their fitness evaluated as well:

Individual 1 : 100101 | 1011 Child 13 : 1001010111 Fitness : 20%
Individual 3 : 100010 | 0111 Child 31 : 1000101011 Fitness : 67%
- - - - - - - - - - - - - - - - - - -
Individual 2 : 0101 | 110110 Child 24 : 0101000011 Fitness : 80%
Individual 4 : 1011 | 000011 Child 42 : 1011010110 Fitness : 40%

Once the new children have been evaluated, the top candidates can replace the lowest rank candidates from the original population. The Genetic Algorithm should converge to at least a local minimum, that being the main complaint against the algorithm. Many alternative implementations have been formulated in order to help the GA to converge to an optimal solution. The wrapper feature selection algorithm should converge to the highest classification accuracy possible for the given feature set given a high initial population and number of generations.

GA Feature Selection Results

Several wrappers were constructed using alternative settings to compare the outcome of feature selection. The wrapper was constructed not only with the J48 decision tree, but also with an artificial neural network classifier.

Training time for these feature selection methods varies depending on the classifier and the number of cross-validation folds that will be used to train the classifier based on the selected features. The initial setting for the population size is also a large

determining factor. An efficient implementation would not evaluate children derived from a parent population that is present in the parent population and has already been evaluated. At maximum, each new generation would have as many new children to evaluate as the parent population.

Table 7: GA neural network result

Number of folds	(%)		Attribute
2	(100%)	1	AccessType
1	(50%)	2	TimeOnDialysis
1	(50%)	3	InfedDose
2	(100%)	4	FerritinLabResult
2	(100%)	5	MCVLabResult
2	(100%)	6	AlbuminLabResult
2	(100%)	7	WBCLabResult
0	(0%)	8	StartSittingBP_systolic
1	(50%)	9	StartSittingBP_diastolic
0	(0%)	10	StartSittingBP_difference
0	(0%)	11	EndSittingBP_systolic
1	(50%)	12	EndSittingBP_diastolic
0	(0%)	13	EndSittingBP_difference
2	(100%)	14	EPO/KG
2	(100%)	15	Gender
2	(100%)	16	Ethnicity
2	(100%)	17	Race

The results of Table 7 are interpreted in the following manner: any attribute that was utilized for any number of folds greater than zero was selected by the GA wrapper. This wrapper using an ANN for its subset evaluator eliminated several blood pressure measures, both before and after the dialysis run. All other measures were evaluated as being significant. This feature subset (14 attributes) was then evaluated for classification accuracy using the ANN and the J48 (the GA wrapper does not report the final fitness

values). Final accuracies for the J48 and ANN after feature subset reduction was 77.74% and 53.58%.

Table 8: Results of the GA J48 wrapper

Number of folds	(%)		Attribute
0	(0%)	1	AccessType
0	(0%)	2	TimeOnDialysis
1	(50%)	3	InfedDose
2	(100%)	4	FerritinLabResult
2	(100%)	5	MCVLabResult
2	(100%)	6	AlbuminLabResult
2	(100%)	7	WBCLabResult
0	(0%)	8	StartSittingBP_systolic
0	(0%)	9	StartSittingBP_diastolic
0	(0%)	10	StartSittingBP_difference
1	(50%)	11	EndSittingBP_systolic
0	(0%)	12	EndSittingBP_diastolic
0	(0%)	13	EndSittingBP_difference
2	(100%)	14	EPO/KG
2	(100%)	15	Gender
1	(50%)	16	Ethnicity
0	(0%)	17	Race

The J48 wrapper and the ANN wrapper both selected the following attributes: InfedDose, FerritinLabResult, MCVLabResult, AlbuminLabResult, WBCLabResult, EPO/KG, Gender, and Ethnicity. This result is encouraging given the clinical knowledge used to derive the baseline dataset. Blood pressures are typically more random, and therefore not necessarily a good predictor. The accuracies of the J48 and ANN were determined to be 76.40% and 44.10%, respectively. Though the accuracies for both datasets are very similar, J48 selected variables will be selected based on the simplicity of the model. This will also help to reduce training time.

Classification Experimentation

Data Preprocessing

Data preprocessing is an extremely important step in any data mining experiment. The following sections outline some of the major tasks.

Handling Missing Values

Several strategies are employed to handle missing data for the kidney dialysis patients. Parameters such as weight and blood pressure were estimated using an average of the previous twenty (if available) dialysis runs. The patient's weight would typically fluctuate very little from session to session, whereas the blood pressures typically take on a value less correlated with previous data and appear much more random.

Medication dosages are associated with a particular dialysis run identification number. Therefore, if no medication dosage is found for a particular run ID, a zero dosage is associated with that run. Unfortunately, recorded medication administrations did not record the value of the drug being administered, simply that the drug was administered. These runs were eliminated from the dataset due to the fact that there is no certainty in estimating any value for these instances.

Records for lab values (such as hemoglobin and white blood cell counts) had far fewer data entry errors than the medications administered. The major obstacle for these data fields were large periods of time were not available in the data base. The omissions for the features were not mutually exclusive – if one value was missing, it was highly likely that all lab values were unable for that time period. These runs were eliminated due to the high amounts of estimating that would be necessary to complete the patient's record.

Hemoglobin Discretization

The majority of classification algorithms require that the decision variable be categorical in nature. The size of the bins and the number of examples associated with the bins can affect the ultimate accuracy of the model. To illustrate, the validation set was used with different levels of discretization to study the classification accuracy of the J48 decision tree algorithm.

Table 9: Discretization experiment illustrates affects of bin size

Discretization Level	Classification Accuracy
0.5 g/dL	74.47%
1.0 g/dL	81.18%
1.5 g/dL	86.22%
{Low, Normal, High}	85.99%

As the granularity of discretization decreases (i.e. smaller number of bins), the classification accuracy increases. While this increase in accuracy is desirable, in practice models and rules generated from data mining should help clinicians move a patient from one adjacent state to the next. The differences between patients at adjacent levels should be fewer when the granularity in outcomes is highest.

Validation Set – Classification Accuracy Experiments

Utilizing the features from

Table 6 and data from twenty two randomly selected patients, classification models of the hemoglobin test result were constructed. Several algorithms were chosen

to evaluate their ability to correctly classify the test outcome. A support vector machine classifier was not used because of the categorical nature of several of the features. The test result outcome was discretized to intervals of 0.5 g/dl from 8.5 to 14 g/dl.

Table 10: Classification Accuracy of Various Classifiers

Classifier	Classification Accuracy
Artificial Neural Network	53.71%
C4.5 Decision Tree	76.24%
PART algorithm	56.74%
Naïve Bayes	19.66%

The PART algorithm utilizes the C4.5 algorithm in building a rule base, but there are differences between PART and J48 (a Java working of an updated C4.5) in terms of pruning. Table 11 illustrates the confusion matrix for the PART algorithm results from the validation set. Cells highlighted in red are perfectly predicted by the algorithm. Cells highlighted in orange are what could be considered "neighborhood accuracy". This term essentially means that if we consider the adjacent categories in an ordinal list to be similar, a misclassification to one of these adjacent bins can be considered "in the neighborhood" of accuracy. The ability to draw inference from neighborhood accuracy must be defined by the categories themselves – if the split points between categories are not significant, then we can consider neighborhood accuracy to be important.

An additional aspect of analyzing confusion matrices is how well a particular classifier predicts certain classes. This particular algorithm performs slightly better than a coin toss overall – but does it predict certain classes better than others? Classes that do not contain many examples are typically predicted worse than classes that have many examples. This certainly holds true for the lowest and highest classes in the validation

set. Lack of data for the high categories makes clinical sense – when patients approach a high level of hemoglobin, administration of EPO is scaled back or not administered at all.

Table 11: PART Algorithm confusion matrix

a	b	c	d	e	f	g	h	i	j	k	l	m	<--	classified as		
0	0	1	12	1	0	0	0	0	0	0	0	0	|	a	=	LT_8.5
0	0	2	0	1	0	0	2	1	0	0	0	0	|	b	=	8.5_9
0	0	32	0	7	3	0	6	1	2	0	0	0	|	c	=	9_9.5
0	0	7	104	8	28	17	4	5	12	2	0	0	|	d	=	9.5_10
0	0	11	10	84	37	24	8	9	6	2	0	0	|	e	=	10_10.5
0	0	0	12	30	297	34	33	33	8	1	0	0	|	f	=	10.5_11
0	0	2	16	16	63	279	76	51	17	3	3	0	|	g	=	11_11.5
0	0	1	7	13	53	71	368	80	32	9	3	0	|	h	=	11.5_12
0	0	0	4	9	23	40	89	435	64	14	0	0	|	i	=	12_12.5
0	0	0	1	4	19	26	55	85	355	27	8	10	|	j	=	12.5_13
0	0	0	0	2	6	7	20	36	34	99	4	0	|	k	=	13_13.5
0	0	0	0	2	3	2	11	19	17	16	15	5	|	l	=	13.5_14
0	0	0	1	1	1	2	0	0	21	3	1	15	|	m	=	HT_14

This guideline is set forth by many insuring agencies – they are unwilling to pay for treatment above and beyond the target range of hemoglobin. As for the lowest categories, the body produces erythropoietin in areas other than the kidneys. Therefore, at extremely low kidney function, there will be a natural floor for the body's production of erythropoietin and therefore hemoglobin would not fall below a certain level under normal circumstances.

Artificial neural networks (ANN) have been popular in many fields for a number of years. They have the unique distinction of being known as a "universal approximator" – having the ability to model any type of function. ANNs mimic the neural pathways of the human brain and are trained by a method called error back-propagation. ANNs are mostly described as a "black box" algorithm – the form of the model is difficult to discern. Researchers concerned with the overall form of the model describing a process

should shy away from this form of prediction. Data mining is typically more concerned with outcomes, and having such a large number of tools at disposal, the final algorithm chosen can be affected by personal preference or the characteristics of the dataset.

Naïve Bayes classifiers use Bayes's rule relating joint probabilities to conditional probabilities, along with the conditional independence of individual features, to build a classification model. Performance of such classifiers is degraded by correlated features that violate the conditional independence assumption. Also, continuous features must either be discretized or assumed to follow a probability distribution – typically assumed to be Gaussian in most implementations. Many attributes from the data set would be nicely approximated by a Normal distribution – patient blood pressure or dialysate temperature variables are good examples. Lab results and medication doses may not follow such a distribution. These assumptions typically lead the Naïve Bayes classifier to perform poorly in relation to other classifiers such as neural networks and decision trees that need not make such assumptions about the structure of the source data.

PART Rule Evaluation

Domain knowledge is essential to validate the validity of rules derived from the sample population. What can be defined as a "good" rule can vary, but typically a rule that covers a significant portion of the data instances are most suitable for evaluation. Many rules are typically generated that may only apply to a handful of instances (between 1 and 10). Some of the interesting rules generated are presented next. All are selected on the basis of the number of instances in the dataset they cover. The rules represent the highest coverage of the validation set population.

- Rule 1: IF AlbuminLabResult > 4.1 AND TimeOnDialysis <= 269 AND FerritinLabResult <= 150: High (107.0/1.0)
- Rule 2: IF AlbuminLabResult > 4.1 AND EPO/KG > 43.2 AND WBCLabResult <= 7.8: Low (15.0)

- Rule 3: IF MCVLabResult > 85.599998 AND AlbuminLabResult > 4.1 AND AccessType = TessioCatheter: High (65.0)
- Rule 4: IF AlbuminLabResult > 4.2 AND Ethnicity = 1: Normal (15.0)

The emergence of time on dialysis in Rule 1, and specifically at such a break point, appears to be a little misleading. In practice, most patients are dialyzed for approximately three hours, making a break point of 270 minutes (4.5 hours) meaningless. One would be hard pressed to find many patients within the entire dataset that received treatment for over this length of time. This is just one example of the spurious connections that can be made when implementing data mining. Other rules regarding the split values of lab results or medications require domain expertise in order to fully evaluate them for clinical validity.

Medication Dose Response: Time Series Analysis

Medications are delivered to patients in order to elicit some sort of positive biological response. An increased dose of erythropoietin should elicit a positive change in hemoglobin (holding all other factors constant) some time in the future. This receptiveness can be defined as a dose-response curve – the amount of time required to see a target change in hemoglobin based on a change in erythropoietin dose. For typical patient, hemoglobin is measured approximately every two weeks. The fluctuation between measurements for each patient is unknown, but is highly patient dependent. Patients will have varying degrees of renal function producing varying quantities of erythropoietin.

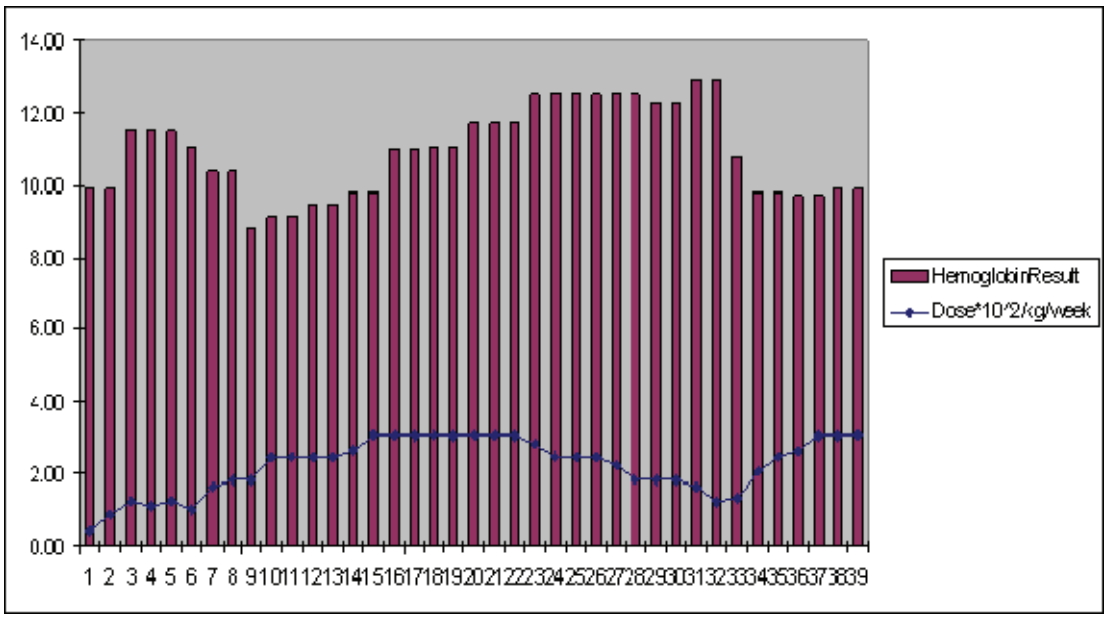

Figure 14: Scaled Epogen dose (units/kg/week) and Hemoglobin for a single patient

Figure 14 shows the treatment of a single patient over the course of thirty-nine weeks. Hemoglobin is held constant until a new value is determined through laboratory results. The target range for most patients is between 11 g/dl and 12 g/dl. As the graph indicates, this value is rarely maintained for long. As hemoglobin approaches 13 g/dl, erythropoietin is no longer administered in some patients. This approach is supported by Phrommintikul et. al.: higher hemoglobin levels are associated with all-causes mortality. Drastic changes in hemoglobin can be attributed to blood loss, hospitalizations, infections, or inflammation. As hemoglobin peaks in Figure 14, erythropoietin dose is being decreased.

Time-series analysis is a very important field of research, and there are many approaches to dealing with this type of data. Temporal data mining enhances this field past simple statistical analysis to determine the optimal set of features to predict some value given past states.

Allen (1984) is an oft referenced paper dealing with the abstraction of temporal events using first-order logical statements. Temporal data mining researchers use these abstractions to define regions or states in time-series data. A direct application in this field is the research conducted by Bellazzi *et al.* [24] with quality assessment of hemodialysis services. Their research involved vast preprocessing using reduction methods, multi-scale filtering, along with temporal abstractions. They were able to define Apriori association rules making extensive use of the temporal operator PRECEDES to identify scenarios that predicted the emergence of a failure event. The resulting rules were an affirmation of clinical domain knowledge, but also generated some unexpected results, which is typical of many data mining applications.

Dynamic control modeling uses techniques known as "system identification" to determine mathematical models that describe a process [22]. The dynamic system is identified by a dataset with an output *y(t)* at some time *t* and also with inputs defined by *u(t)*.

$$Z^t = \{y(1), u(1), ..., y(t), u(t)\}$$

Espinosa et al. [22] state that a model can be formulated to predict the next output using historical data. This model would take the following form.

$$\hat{y}(t) = f(Z^{t-1})$$

This simple function can be defined in any number of ways using any number of algorithms or functions. Data mining algorithms would lend themselves quite well to training these functions. The authors propose a set of pertinent questions, including how to choose the predictors and the optimal time delay for this function. Fuzzy logic, correlation, or other statistical methods (like principal component analysis) are typical methods of choosing predictors. Feature selection searches (such as greedy, entropy-

based searches or genetic algorithm wrapper methods) can find the optimal combination of predictors.

The system identification approach can be applied to modeling the response of a patient to hemoglobin treatment. As an example of how this would be accomplished, consider a feature set that contains the weekly dosage of erythropoietin and the latest result of the patient's hemoglobin test. A delay of five weeks is arbitrarily defined for analysis. A sampling of the constructed dataset is shown below.

Table 12: Predictor data for current hemoglobin using previous weeks data

Hgb-1	Hgb-2	Hgb-3	Hgb-4	Hgb-5	EPO	EPO-1	EPO-2	EPO-3	EPO-4	EPO-5	Curr Hgb
11.50	11.50	11.50	9.90	9.90	10000	12200	10800	12000	8400	4200	11.10
11.10	11.50	11.50	11.50	9.90	16000	10000	12200	10800	12000	8400	10.40
10.40	11.10	11.50	11.50	11.50	18000	16000	10000	12200	10800	12000	10.40
10.40	10.40	11.10	11.50	11.50	18000	18000	16000	10000	12200	10800	8.80
8.80	10.40	10.40	11.10	11.50	24000	18000	18000	16000	10000	12200	9.10
9.10	8.80	10.40	10.40	11.10	24000	24000	18000	18000	16000	10000	9.10

The Epogen dosages are the summation of the dosages administered during the dialysis runs that occur during a weekly basis. This summation may be more comfortable for clinicians considering that is how the dose is typically prescribed. "Hgb-1" refers to the previous weeks Hgb lab value, and a similar notation is used for the previous five weeks of both Epogen and Hgb. This data comes from a single patient, and there were over 40 weeks of data available where Epogen was administered. The outcome "CurrHgb" was discretized to the categories "Low, Normal, High" as before because of the small size of the data set, and so the J48 algorithm could be utilized.

Can we accurately predict the current value of hemoglobin given previous hemoglobin information and erythropoietin dose? At first glance, and knowing how the dataset was constructed, the previous week's Hgb value would be intuitively selected to

predict the current state of Hgb. The majority of previous values will be the same as the current values, and there are typically no wild swings in Hgb unless some other traumatic event has occurred (i.e. blood loss). Which parameters will the J48 algorithm select as accurate predictors of the patient's Hgb status?

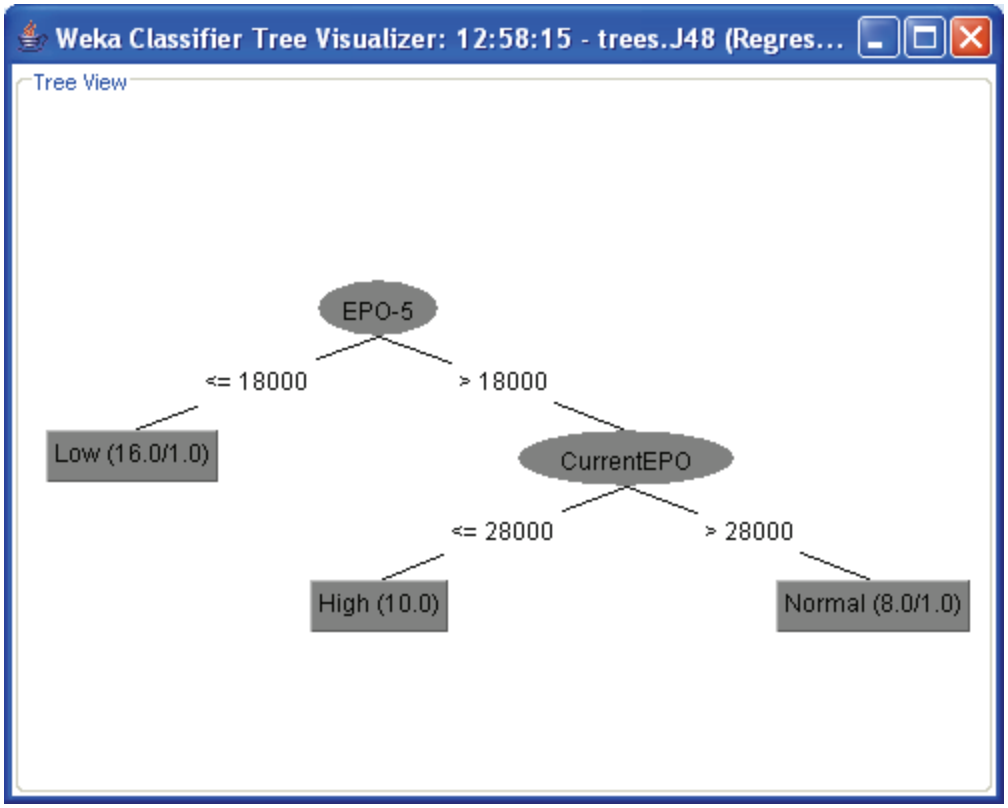

Figure 15: Time-series predicator analysis with J48 decision tree

This very simple tree, though not generalizable to the entire population, does provide a very interesting result. In this particular patient with this particular discretization of Hgb, the only important predictors of the current Hgb status were the EPO dose from 5 weeks ago and the current EPO dosage. A simple transformation of the decision tree results in the following decision rules:

Rule 1: IF EPO-5 <= 18000 THEN CurrentHgb = Low (16.0/1.0)

Rule 2: IF EPO-5 > 18000 AND CurrentEPO <= 28000 THEN CurrentHgb = High (10.0/0.0)

Rule 3: IF EPO-5 > 18000 AND CurrentEPO > 28000 THEN CurrentHgb = Normal (8.0/1.0)

The values in the parentheses are the correctly classified and misclassified instances. For Rule 1 there are 16 instances correctly classified and 1 incorrectly classified.

Transferring these rules into clinical knowledge for this patient would indicate that if in 5 weeks the desired Hgb will fall in a "Normal" or above range, the patient should be given at least 18,000 units per week of EPO in order to achieve this goal, all else being equal. Split points for each attribute are determined by using the entropy calculation over a number of potential split points and greedily selecting the best choice. Looking at the data, the EPO-5 doses that were associated with normal or high Hgb were typically much higher than 18000 units (see Table 12)

Being able to define the predictive function *f* using data mining has distinct advantages over current approaches. Most notably this function can be retrained in order to more accurately reflect a systems current behavior. Another benefit arises from the inherent ability to accurately learn and describe complex systems. Using system identification in such a way allows you to formulate the problem mathematically in order to optimize a specific future outcome. This optimization can be formulated in such a way as to explore the unknown regions of the current control (treatment) space. Padmanabhan and Tuzhilin [25] have demonstrated how data mining can be used to specify the three most important aspects of any optimization problem: the state space, objective function, and the system constraints. Being able to use data mining in such a way also provides a framework to find a proper treatment to elicit a desired outcome (i.e. drug administration to control anemia)

CLUSTERING-DERIVED POPULATIONS – CLASSIFICATION IMPLICATIONS

What is the ultimate affect of collecting the dialysis run information and building classifiers based on a cluster assignment of a patient? If these patients are considered similar by a clustering algorithm, will decision trees and rules be more or less accurate? Will the clustering increase the coverage of derived rules, making them an overall better predictor of clinical behavior/patient response? How will the prediction accuracy compare to randomly generated patient datasets.

Having created a collection of patients using clustering that can now be considered a homogeneous population, predictive modeling techniques may be implemented more accurately. Taking the collection of time series data for all the patients within a specific subgroup, we can explore the various factors leading to the onset of anemia in the dialysis patients. There are many data mining algorithms that feature rule sets as output including association rules and decision tree algorithms. Algorithms featuring rule set generation may be preferred over other techniques, such as neural networks, that do not produce such an output.

Related Literature

This methodology of employing clustering to enhance the classifications made by decision trees was employed by Gaddam *et al*. [14] in an anomaly detection domain. Their study employed K-Means clustering to group objects into normal or anomalous classes in multiple domains (computer network traffic, electronic circuitry, and a mechanical mass-beam system). The ID3 decision tree algorithm is then employed to further understand the particular patterns within the normal and anomalous classes.

Pedrycz and Sosnowski [17] implemented a fuzzy c-means algorithm to grow decision trees with a radical new geometry. These non-linear trees consider heterogeneous subsets of data to build diverse C-decision trees. In contrast to more

common implementations, however, the C-fuzzy decision trees consider many features at a node as opposed to the single node considered in C4.5.

Common medical studies resemble the research paper by Hernándex-Garduño *et al*. [20]: identifying a group of patients according to a test value or procedure and evaluating the risk factors associated with a condition. In this case, the medical team clustered patients using the identification of DNA patterns in Mycobacterium tuberculosis. Using DNA pattern recognition software, patients were established as either having the disease or not. Combining principal component analysis with multivariate logistic regression, the researchers were able to determine the strongest predictors of having the condition. An interesting extension of this study would be to determine how well data mining algorithms such as a decision tree would predict the cluster assignment.

Experimentation

The clustering accomplished in the first section and the decision tree analyses from section two are combined. Decision trees are constructed based on the cluster assignment from k-means clustering and EM algorithm clustering. If clustered patients are to be considered similar, decision trees constructed based on run information should be more accurate than data sets constructed from random samples.

Clusters were not constructed on the basis that the patients were of similar hemoglobin distribution. Nor were patients clustered on the basis of distribution of any medication dosages.

Section two established the supremacy of classifiers based on the C4.5 algorithm in terms of classification accuracy for the validation set. These classifiers did consider the administration of the important drug erythropoietin, which was confirmed to be of importance using both the C4.5 decision tree and the multi-layer perceptron neural network genetic algorithm wrappers. To decrease training time, the feature subset used

to train the clustered classification models was derived from the two-fold decision tree selection. While C4.5 classifiers performed most accurately, the validation set was too small to derive any general conclusions. Therefore, cross-validation accuracy was evaluated for all K-Means clusters using J48 and ANN. The PART algorithm was also used to verify any discrepancies in implementation.

Table 13: 10-fold cross validation accuracy for various algorithms

Dataset	J48	ANN	PART	J48 Top Node	Instances
KMeansCluster0	95.99%	55.021%	96.03%	Ethnicity	14051
KMeansCluster1	83.63%	53.13%	83.58%	WBC	11711
KMeansCluster2	77.00%	26.43%	76.91%	Ferritin	7910
KMeansCluster3	88.73%	27.26%	88.67%	Alb	12059
KMeansCluster4	78.33%	37.83%	77.72%	WBC	4713
KMeansCluster5	73.27%	27.44%	73.33%	Alb	12263
KMeansCluster6	74.32%	30.52%	74.48%	Alb	11526
KMeansCluster7	68.89%	32.28%	69.08%	Alb	5641
KMeansCluster8	84.67%	72.15%	84.49%	Alb	1070
KMeansCluster9	86.11%	22.96%	85.71%	Alb	18757

As indicated in Table 13, the C4.5 algorithm performs much better than the artificial neural network. At maximum, the J48 algorithm (based on C4.5) classifies the hemoglobin of cluster 0 patients at nearly 96% accuracy. The model performs the most poorly on the data derived from patients determined to be in cluster 7 – 69% accurate. Accuracy for the PART algorithm should and is extremely closely correlated to the accuracy for the J48 algorithm. The neural network performs best for cluster 8 (72.15%) and worst for cluster 9 (23%).

An important feature of the J48 algorithm, particularly for clinicians, is which parameter was chosen to be the most important parameter to initially split the data. In general, Albumin was selected as the most important split point for the entire dataset.

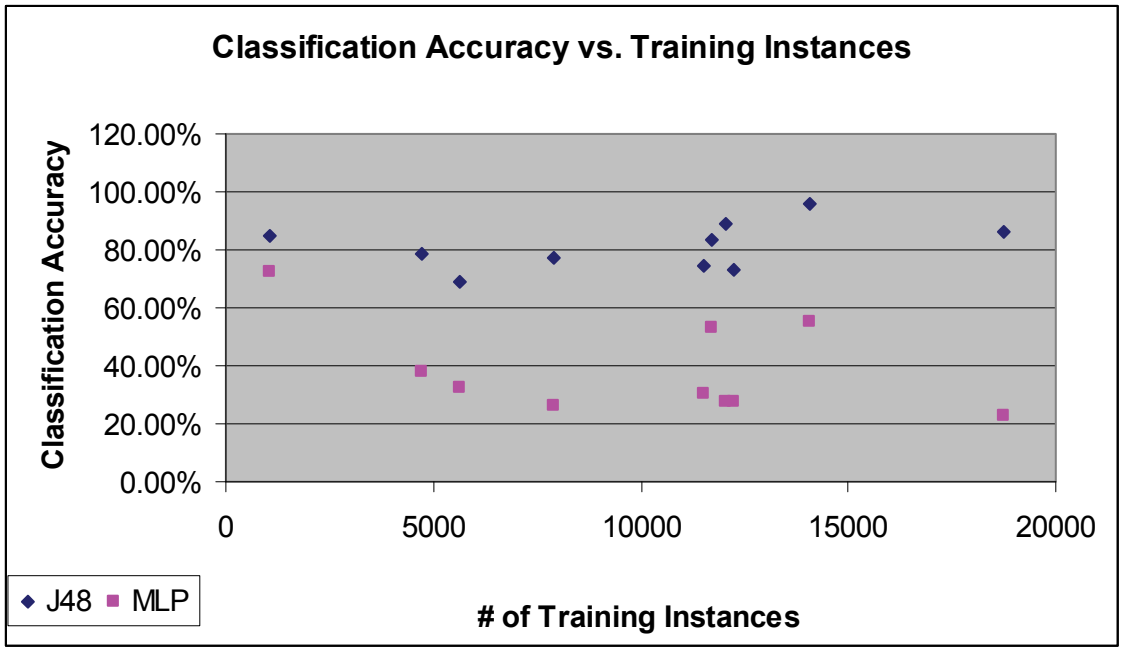

Figure 16: Classification Accuracy is not affected by the number of training instances

There is not a noticeable increase in classification accuracy among the datasets as the number of training instances increase, as is sometimes expected. However, larger datasets will typically generalize better to unseen testing instances. This hypothesis can be tested on the validation set, which follows in the next section.

Training of the EM generated clusters used only the J48 algorithm and ANN, as the PART algorithm was (and should) be highly correlated with the J48 output. The results of the 10-fold cross-validation follow. Cluster 2 was eliminated from analysis due to a lack of data for many of the laboratory values, including hemoglobin.

Table 14: EM generated clusters - Classification Accuracy

Dataset	J48	MLP	J48 Top Node	Instances
EMCluster0	89.06%	22.08%	Alb	22317
EMCluster1	78.71%	21.61%	Alb	31395
EMCluster3	72.58%	24.61%	Alb	29319
EMCluster4	77.45%	27.88%	Ferritin	9931
EMCluster5	91.86%	33.29%	EPO/KG	5578

The classification accuracy for the J48 algorithm is very comparable to the K-Means clusters. The accuracy for the ANN is still very poor, but makes intuitive sense considering the use of the J48 selected feature set. The role of the number of training instances does not appear to have much effect again with the EM clusters. Interestingly, cluster 5 produced the highest cross-validation accuracy with the fewest training instances. This dataset also selected the Epogen dosage as its most important predictor to initially split the dataset.

Testing Accuracy – Validation Sets

The validation set consists of information from 21 randomly selected patients. 20 of these patients were included in the clustering dataset. The cluster assignment for these patients is indicated in the table below.

Table 15: Cluster assignments for validation set members

PatientID	KMeansCluster	EMCluster
89007	cluster0	cluster1
242023	cluster9	cluster2
247022	cluster4	cluster4
252023	cluster9	cluster0
296024	cluster1	cluster5
380023	cluster3	cluster0
476001	cluster5	cluster3
492001	cluster6	cluster3
506001	cluster6	cluster3
511001	cluster6	cluster3
529001	cluster6	cluster3
544001	No Cluster	No Cluster
546001	cluster7	cluster3
572001	cluster6	cluster3
577001	cluster6	cluster3
600001	cluster9	cluster1
612001	cluster5	cluster3
672003	cluster9	cluster0
674005	cluster7	cluster3
781015	cluster0	cluster1
961001	cluster6	cluster3

Seven of the 21 patients were included in cluster 6, which should indicate that classification for the validation set should be above normal when compared to the rest of the population. Eleven patients were included in cluster 3 of the EM clusters. To overcome the bias inherent in testing a model with essentially resampled data, testing accuracy will also be observed for a randomly selected K-Means cluster (cluster 7). With these two validation sets, a measure for the generalizability of the models should be observed.

Table 16: Testing Accuracy of validation set and cluster 7

Dataset	Validation Set Accuracy	Cluster 7 Accuracy
KMeansCluster0	19.31%	17.07%
KMeansCluster1	15.80%	10.85%
KMeansCluster2	10.32%	12.50%
KMeansCluster3	14.41%	11.20%
KMeansCluster4	18.31%	13.28%
KMeansCluster5	20.02%	13.19%
KMeansCluster6	41.43%	11.84%
KMeansCluster7	17.24%	83.42%
KMeansCluster8	9.56%	10.81%
KMeansCluster9	20.43%	12.99%

The J48 algorithm was used to evaluate the testing accuracy, and the results are reported in Table 16. The testing accuracies for these two validation sets are very poor. This would appear to infer that constructing datasets according to clustering assignment will lead to a very low generalizability, which is an intuitive outcome. This must be considered in any optimization strategy that includes this type of data sampling.

Being that the EM clusters appear to be slightly more random in composition, they may generalize better than the K-Means clusters. The EM clusters shifted the focus from the categorical parameters to the continuous parameters, which are more randomly distributed in terms of ethnicity, race, and gender.

Table 17: Testing Accuracy for EM Clusters

Dataset	Validation Set Accuracy	Cluster 7 Accuracy
EMCluster0	19.15%	10.12%
EMCluster1	21.14%	13.81%
EMCluster3	48.79%	79.26%
EMCluster4	16.13%	11.65%
EMCluster5	13.29%	11.85%

The EM cluster performed just as poorly on both testing sets. As can be inferred from the results, but proven from the data, EM cluster 3 and K-Means cluster 7 contained many of the same data points. As previously mentioned, half the validation set consisted of patients assigned to EM cluster 3.

Discussion

The poor testing accuracy of the cluster-derived models is not unexpected. In fact this result is quite intuitive – if the cluster populations are to be considered homogeneous, then the models derived should apply most accurately to similar patients. The subsets do appear to be somewhat exclusive in-so-far as not being able to accurately predict the states of either randomly formed subgroups or clustered subgroups. Generalizability could be improved by fully optimizing the feature subset for both clustering and classification applications. As discussed at the end of Chapter 1, these goals are inherently dependent upon each other – the clustering must be subjectively evaluated using the outcomes of the classification, which are in turn dependent on the population derived from the clustering. This could be accomplished iteratively in the form of a search – of which there are many fine choices for methodologies.

CONCLUSION

Chapter 1 determined the usefulness of automated clustering algorithm to detect the presence of clusters of patients using static parameters. Two basic clustering algorithms were implemented – the K-Means algorithm and the Expectation Maximization (EM) algorithm. The clusters for *k* value equal to ten were evaluated for their purity based on the particular features of the datasets. Gender appeared to be a particularly important parameter in determining cluster membership, as the K-Means clusters were typically very pure gender-wise. Factors such as the age of the patient and the number of years that patient had been receiving dialysis were not considered to be as important. Several clusters were very pure in terms of ethnicity and race, while most contained a variety.

The EM algorithm determined that the optimal number of clusters based on the data set to be six. The resulting clusters were far less pure in terms of most of the categorical variables, which are converted automatically to individual features and indicated by a 1 or 0. This conversion is not optimal for these features, as they may become sparse in the conversion process. The six EM clusters were predominately 50% male and female, with other variables having varying distributions.

The inferences that can be drawn from these two vastly different clustering is that the results will depend on a number of factors. The algorithm, feature set, and data types all seem to influence the resulting clusters. The clustering analysis proposed here is by no means exhaustive, as only a limited number of factors were utilized. Future clusterings can be set up in a vast number of configurations, depending on the ultimate goal. Deriving some sort of class label for groups of patients based on their time-series characteristics would be of interest to improve the quality of population-based treatment

Chapter 2 described the processes of deriving a model based on time-series data to predict the current status of a patient's hemoglobin test outcome. Feature subsets were

defined, and subsequently refined to facilitate more efficient training time for classification models. The method employed was a genetic algorithm feature selection wrapper, incorporating both neural networks and J48 decision trees to evaluate potential subsets. The result of this analysis resulted in more compact datasets of comparable cross-validation accuracy.

A key limitation of this study was the lack of some very important information in the data set regarding Epogen dosages over a longer period of time. As a result, many instances of data needed to be eliminated as it would be difficult to estimate these parameters. This is also true of many of the lab values (Hgb, Ferritin, WBC, etc). Feature selection appeared to indicate less of an influence on parameters specifically associated with run information (blood pressure, time on dialysis). Future studies may ignore this data and focus on weekly measurements, which seem to have more applicability to clinicians.

Of vast clinical importance is the ability to forecast a future state of the patient's condition based on current and past information. Temporal data mining techniques and system identification methodologies are introduced. Temporal data mining involves the abstraction of first-order logical concepts to time periods of data. The process proposed by Bellazzi *et al*. [24] involves significant data transformation and denoising in order to apply the abstraction. This method has advantages and disadvantages, but mainly the application of denoising in conjunction with abstractions of real data may reduce the amount of real variability of the process. Depending on the application, this may not be desirable. Patient lab results are typically very stable and highly correlate between periods – an artifact generally ascribed to the fact that some lab values are only collected every three months.

Applying the control theory methodology of system identification appears to have significant clinical merit. Combining this modeling technique with data mining technologies to form a more robust prediction function could advance many areas of

research. This is true particularly in the healthcare domain where treatment outcomes are not immediately known – medications take time to react, the body requires time to recover.

Knowledge derived from a subset of decreasing size but increasing homogeneity is a first step towards treatment individualization. Naturally, the inferences made from these subsets do not transfer as well to related subgroups, as was demonstrated in Chapter 3. As there are a vast number of ways to group patients, there is substantial room for optimization – both in the way patients are grouped and the manner in which the treatment outcome is predicted. However, the capability to use data mining in such a manner has proven that it is worthy of further exploration.

REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. "From Data Mining to Knowledge Discovery in Databases", *American Association for Artificial Intelligence*. 1996.

[2] Berry, M., Linoff, G., "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management", Second Edition 2004.

[3] US Renal Data System: "USRDS 2005 Annual Data Report: Atlas of End-Stage Renal Disease in the United States" Bethesda, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 2005.

[4] Daniel W. Coyne, "Influence of Industry on Renal Guideline Development", *Clinical Journal of the American Society of Nephrology* l 2: 3-7, 2007.

[5] Tilman B. Drüeke, Francesco Locatelli, Naomi Clyne, Kai-Uwe Eckardt, Iain C. Macdougall, Dimitrios Tsakiris, Hans-Ulrich Burger, Armin Scherhag, "Normalization of Hemoglobin Level in Patients with Chronic Kidney Disease and Anemia", *New England Journal of Medicine*, Volume 355:2071-2084.

[6] Ifudu, O., Uribarri, J., Imran Rajwani, Vera Vlacich, Kathy Reydel, Georgina Delosreyes, Eli A. Friedman "Gender modulates responsiveness to recombinant erythropoietin", American Journal of Kidney Diseases, Volume 38, Number 3, 2001.

[7] Besarab A, Bolton WK, Browne JK, Egrie JC, Nissenson AR, Okamoto DM, Schwab SJ, Goodkin DA: "The effects of normal as compared with low hematocrit values in patients with cardiac disease who are receiving hemodialysis and Epoetin" *New England Journal of Medicine* 339:584-590, 1998.

[8] Xia H, Ebben J, Ma JZ, Collins AJ: "Hematocrit levels and hospitalization risks in hemodialysis patients" *Clinical Journal of the American Society of Nephrology* 10:1309-1316, 1999.

[9] Phrommintikul, A., Haas, S., Elsik, M., Krum, H., "Mortality and target haemoglobin concentrations in anaemic patients with chronic kidney disease treated with erythropoietin: a meta-analysis", *Lancet* 2/3/2007, Vol. 369 Issue 9559, p381-388.

[10] Tan, P., Steinback, M., Kumar, V., "Introduction to Data Mining" First Edition 2006.

[11] S. Shah and A. Kusiak, "Cancer Gene Search with Data-Mining and Genetic Algorithms", *Computers in Biology and Medicine*, Vol. 37, No. 2, 2007, pp. 251-261.

[12] Bhamidipati V.R. Murthy, Donald A. Molony, and Austin G. Stack "Survival Advantage of Hispanic Patients Initiating Dialysis in the United States Is Modified by Race" *Clinical Journal of the American Society of Nephrology* 16: 782–790, 2005.

[13] Biagio R. Di Iorio, Davide Stellato, Natale G. De Santo, Massimo Cirillo "Association of Gender and Age with Erythropoietin Resistance in Hemodialysis Patients: Role of Menstrual Status" *Blood Purification* 2004;22:423-427.

[14] Gaddam, S., Phoha, V., Balagani, K., "K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Mean Clustering and ID3 Decision Tree Learning Methods" *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, March 2007.

[15] Albayrak, S., Amasyalı, F., "Fuzzy C-Means Clustering on Medical Diagnostic Systems" *Turkish Symposium on Artificial Intelligence and Neural Networks* 2003.

[16] Weiss, G., Goodnough, L., "Anemia of Chronic Disease" *The New England Journal of Medicine* Volume 352, pp. 1011-23, 2005.

[17] Pedrycz, W., Sosnowski, Z., "C-Fuzzy Decision Trees", *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, Vol. 35, 2005, pp. 498-511.

[18] Ross Quinlan (1993). "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, San Mateo, CA.

[19] Bensmail, H., Meulman, J., "Model-based Clustering with Noise: Bayesian Inference and Estimation" *Journal of Classification* 20: 49-76, 2003.

[20] Hernández-Garduño, E., Kunimoto, D., Wang, L., Rodrigues, M., Elwood, K., Black, W., Mak, S., FritzGerald, J., "Predictors of clustering of tuberculosis in Greater Vancouver: a molecular epidemiologic study", *Canadian Medical Association Journal*, 8/20/2002, Vol. 167, Issue 4, pp. 349-352.

[21] Coulter, JS., "Red cell distribution width and mean corpuscular volume: clinical applications", *Advancing Clinical Care,* 1991 Nov-Dec; 6(6): 13.

[22] Espinosa, J., Vandewalle, J., Wertz, V., "Fuzzy Logic, Identification and Predictive Control" Advances in Industrial Control series, Springer-Verlag London 2005.

[23] Allen, J., "Towards a General Theory of Action and Time" *Artificial Intelligence*, 1984, 23:123-154.

[24] Bellazzi, R., Cristiana, L., Magni, P., Bellazzi, R., "Temporal data mining for the quality assessment of hemodialysis services" *Artificial Intelligence in Medicine*, 2005, 34: 25-39.

[25] Padmanabhan, B., Tuzhilin, A., "On the Use of Optimization for Data Mining: Theoretical Interactions and cCRM Opportunities" *Management Science* Vol. 49, No. 10, October 2003, pp. 1327-1343.

[26] National Kidney and Urologic Diseases Information Clearinghouse, "Anemia in Kidney Disease and Dialysis" NIH Publication No. 05–4619, 2005.

[27] U.S. National Library of Medicine "Anemia" MedlinePlus, access online http://www.nlm.nih.gov/medlineplus/ency/article/000560.htm

[28] University of Virginia Health System "End-Stage Renal Disease (ERSD)" University of Virginia, accessed online, http://www.healthsystem.virginia.edu/uvahealth/adult_urology/endstage.cfm

[30] National Kidney and Urologic Diseases Information Clearinghouse, "Your Kidneys and How They Work" NIH Publication No. 06–4241, 2005.

[31] Blood. (2007). In *Encyclopædia Britannica.* Retrieved March 5, 2007, from Encyclopædia Britannica Online: http://www.britannica.com/eb/article-257804.

[32] Bland, J., "Finding the right therapy: a look at personalized medicine." *Integrative Medicine: A Clinician's Journal,* 5: pp. 10-2, 2006.

[33] Go, V., Wong, D,.Wang, Y., Butrum, R., Norman, H., Wilkerson, L., "Diet and Cancer Prevention: Evidence-based Medicine to Genomic Medicine" *Journal of Nutrition*, Vol. 134, p. 351, 2004.

APPENDIX A: DIALYSIS POPULATION

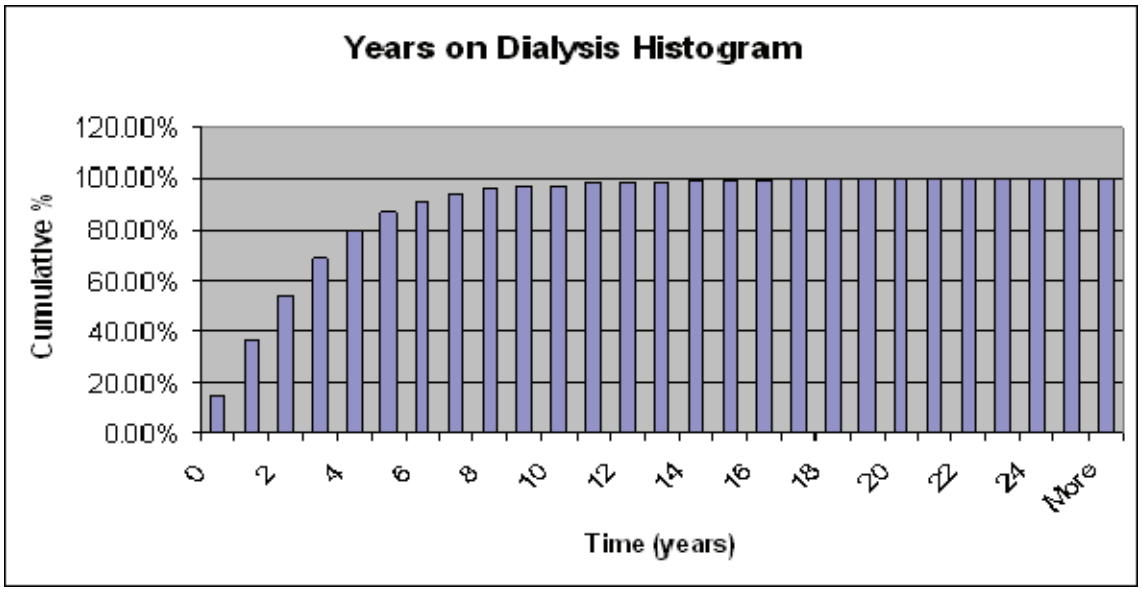

Figure A.1: Cumulative percentage of years on dialysis

APPENDIX B: K-MEANS CLUSTERING DISTRIBUTIONS

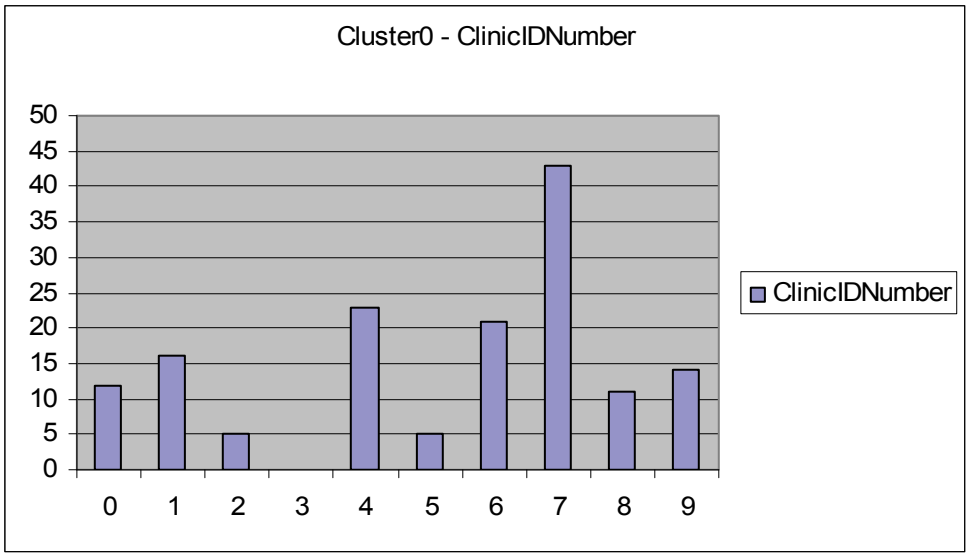

Figure B.1: K-Means Cluster0 ClinicID Number

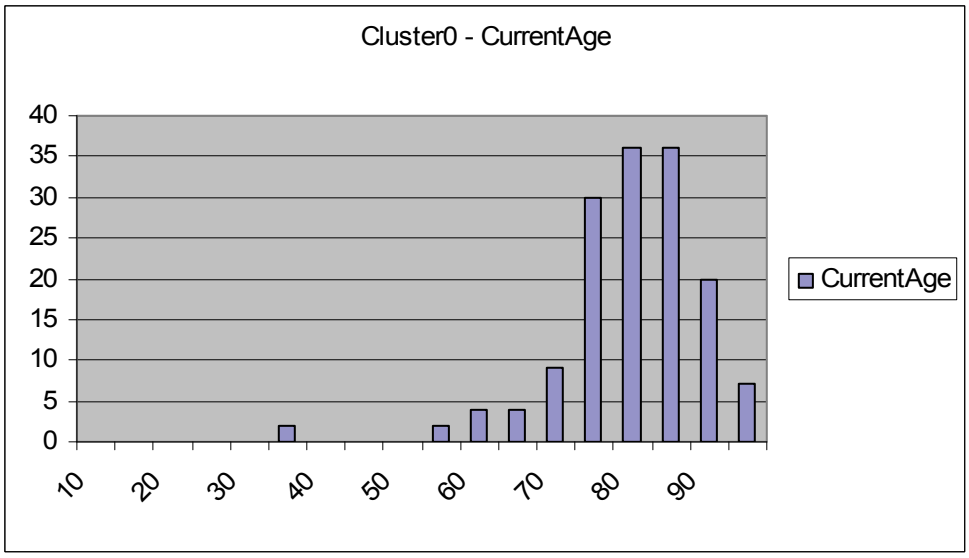

Figure B.2: K-Means Age of Patients in Cluster0

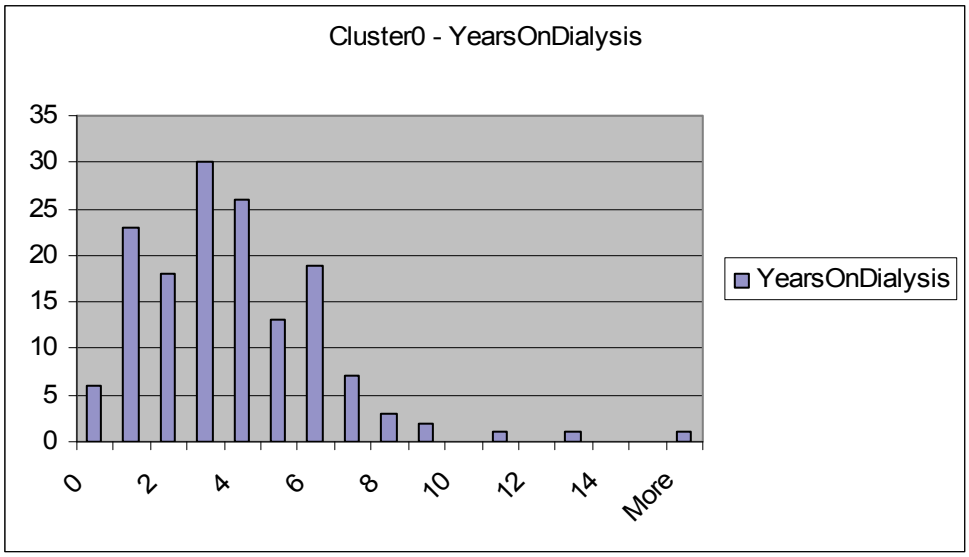

Figure B.3: K-Means Number of Years of Dialysis for Patients in Cluster0

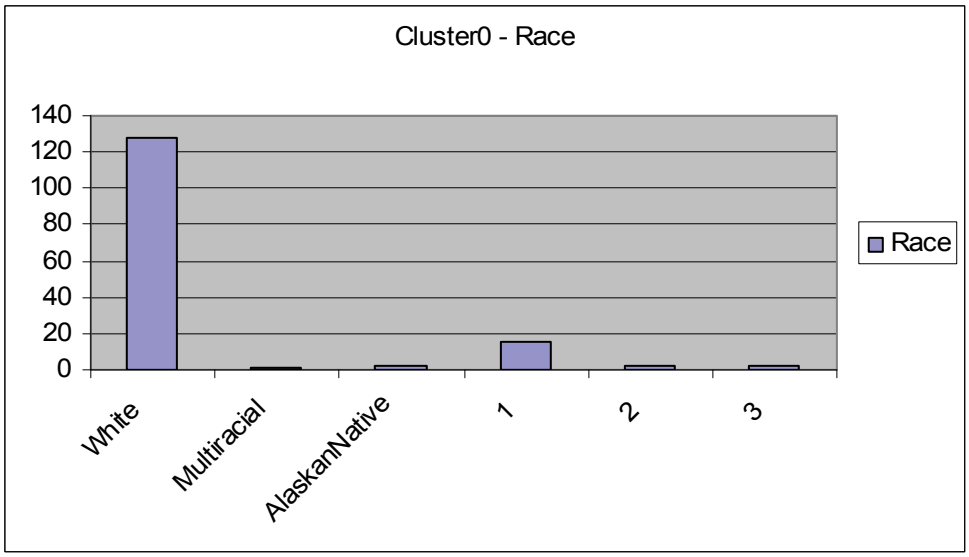

Figure B.4: K-Means Race Distribution from Cluster0

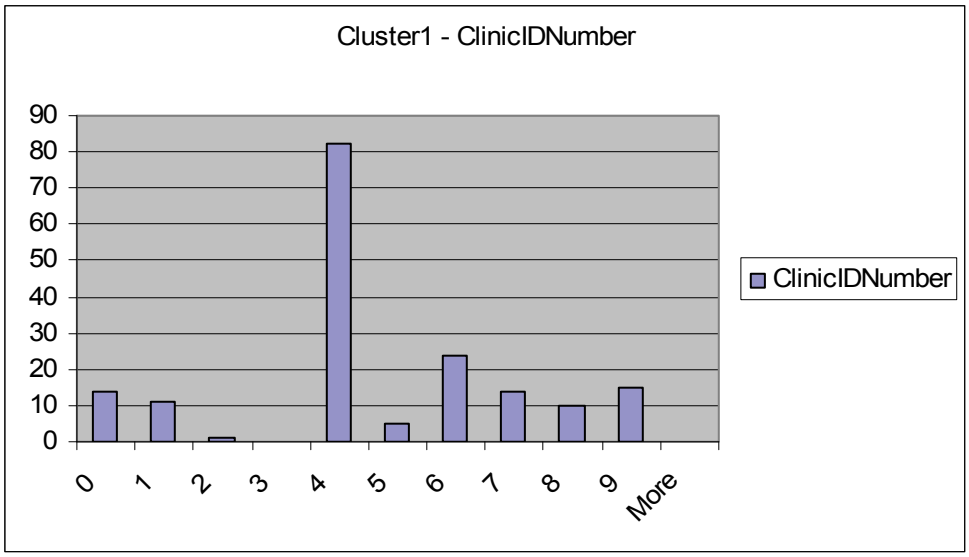

Figure B.5: K-Means Cluster 1 ClinicID Number

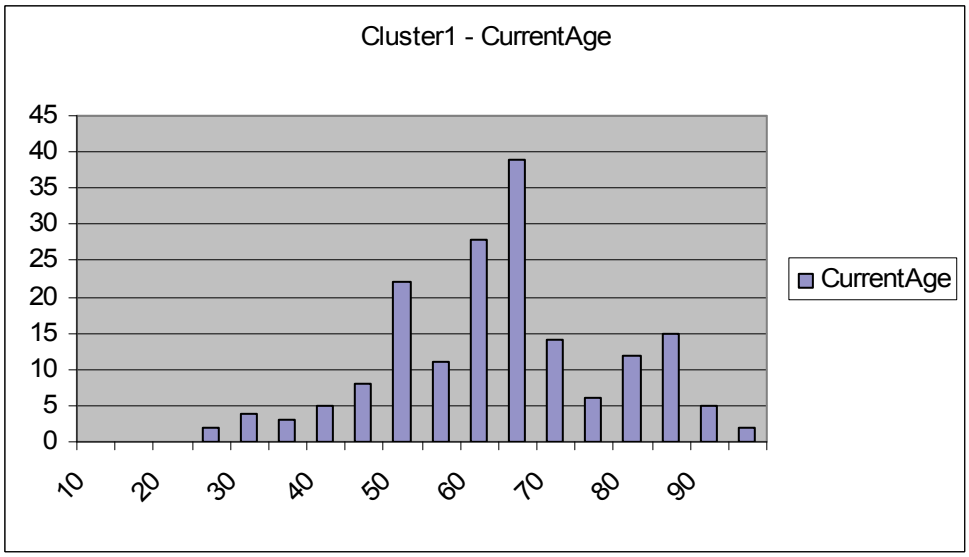

Figure B.6: K-Means Age Distribution of Cluster 1 Patients

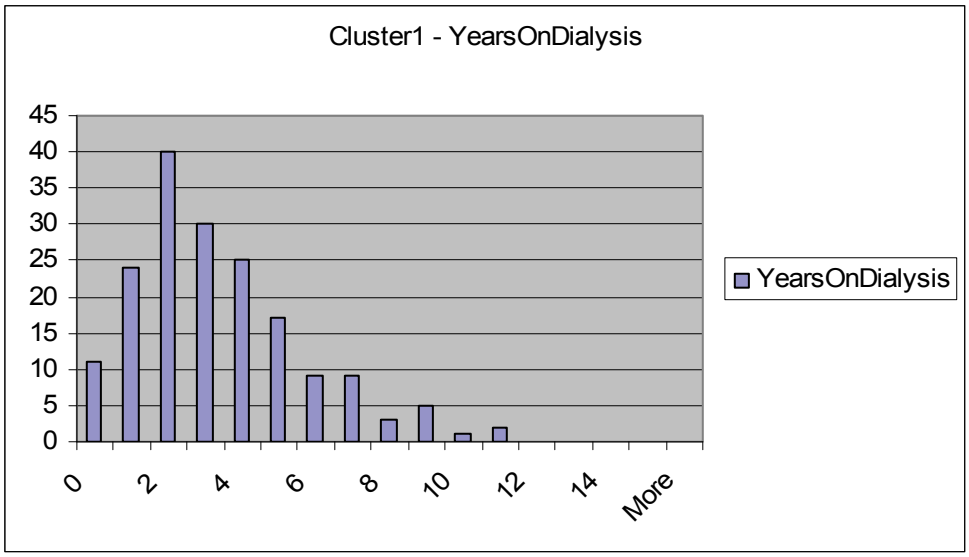

Figure B.7: K-Means Years on Dialysis Distribution for Cluster 1 Patients

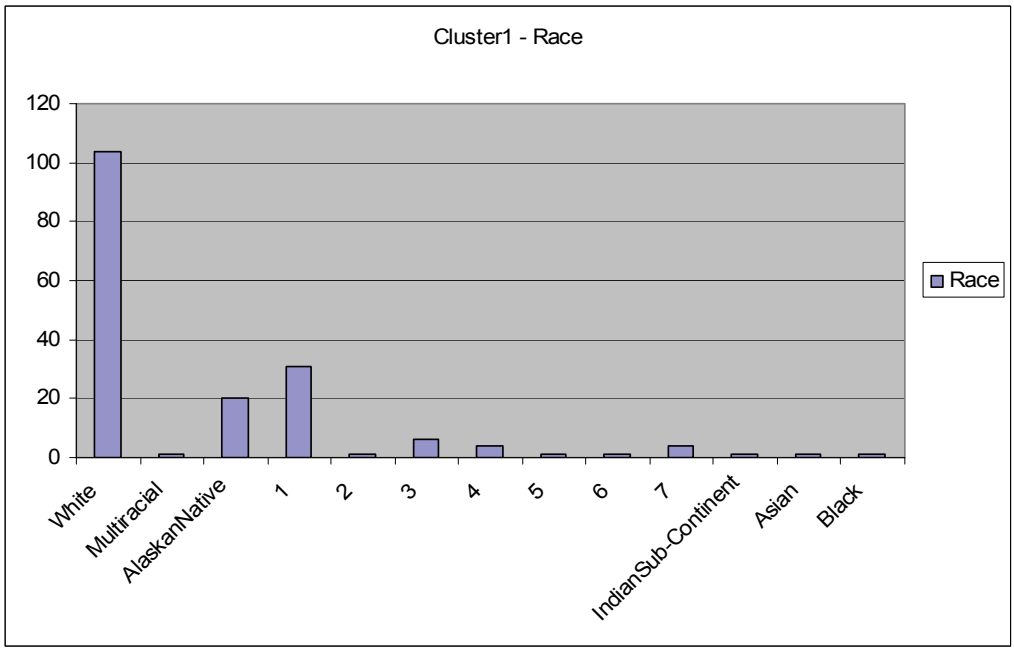

Figure B.8: K-Means Distribution of Races in Cluster 1

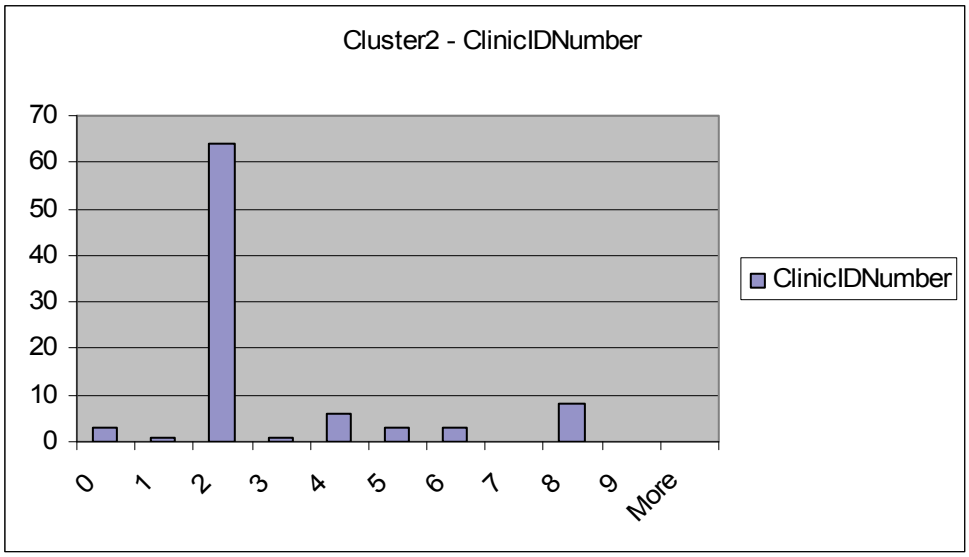

Figure B.9: K-Means Cluster 2 Clinic ID

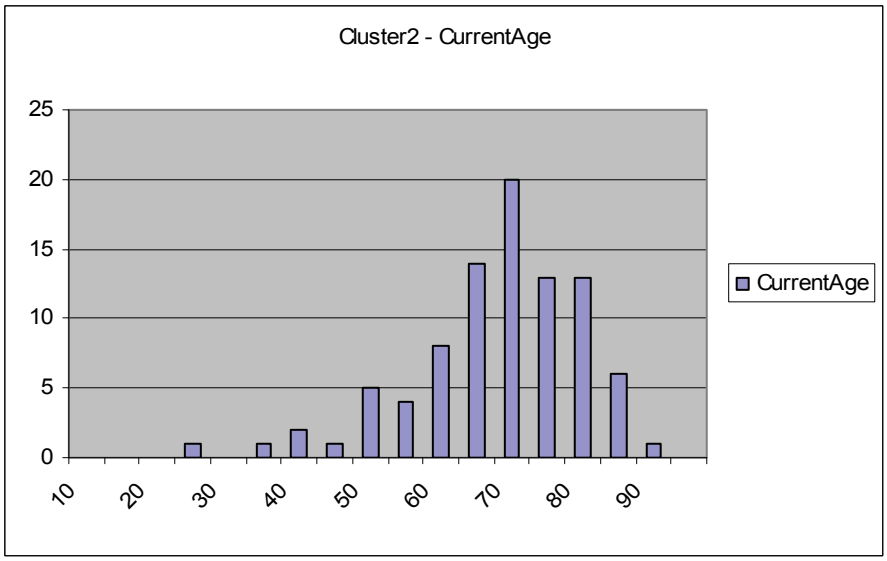

Figure B.10: K-Means Cluster 2 Age Distribution

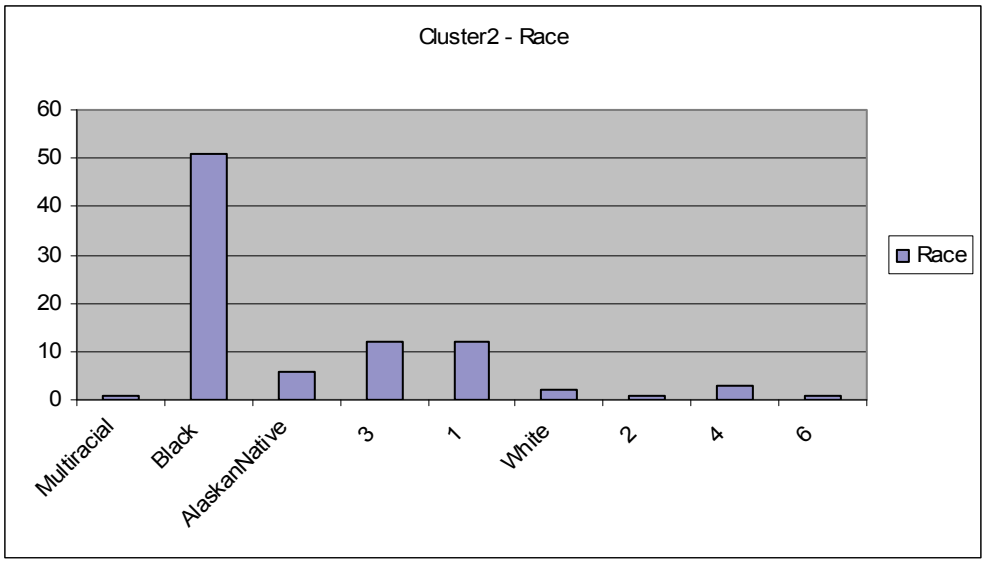

Figure B.11: K-Means Cluster 2 Race Distribution

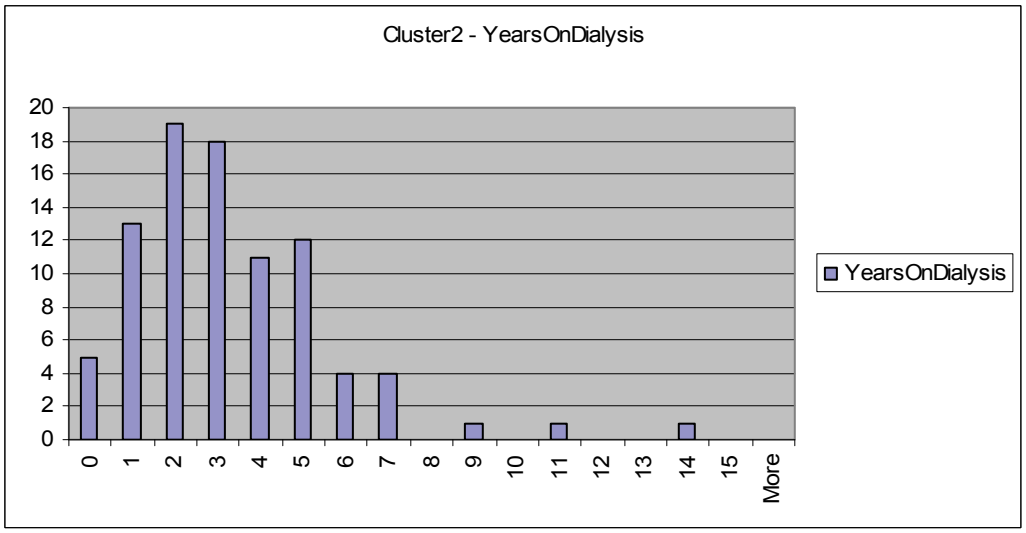

Figure B.12: K-Means Cluster 2 Years on Dialysis Distribution

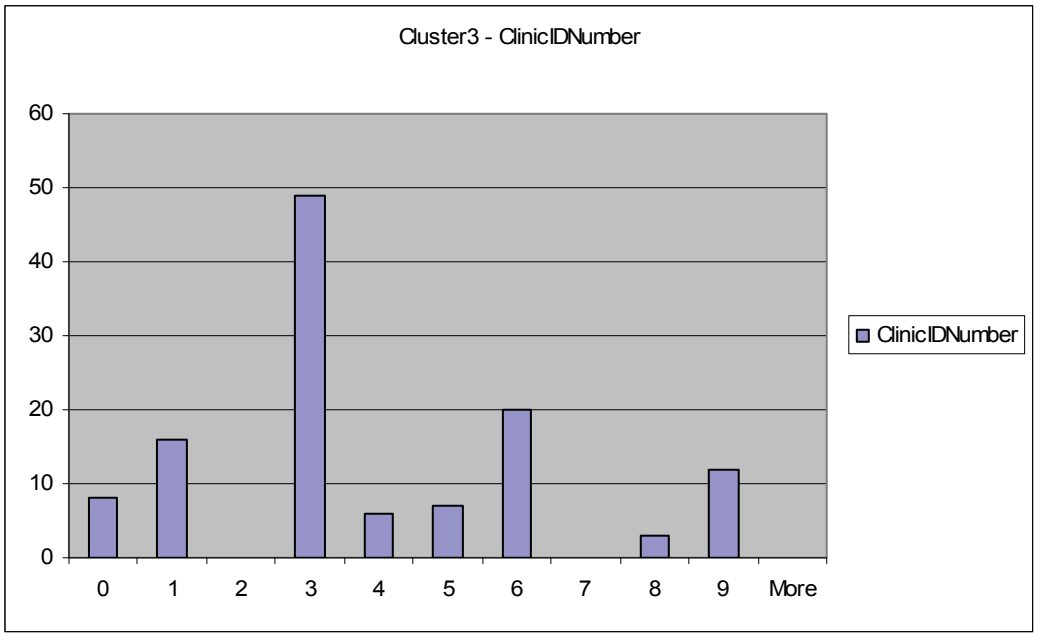

Figure B.13: K-Means Cluster 3 Clinic ID Number

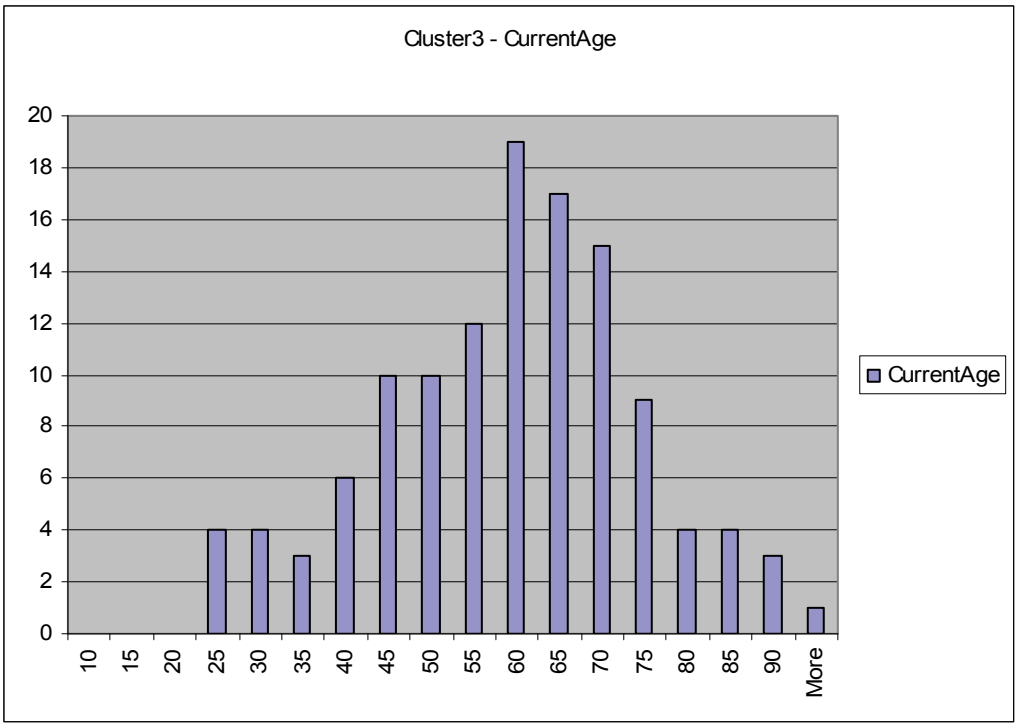

Figure B.14: K-Means Cluster 3 Age Distribution

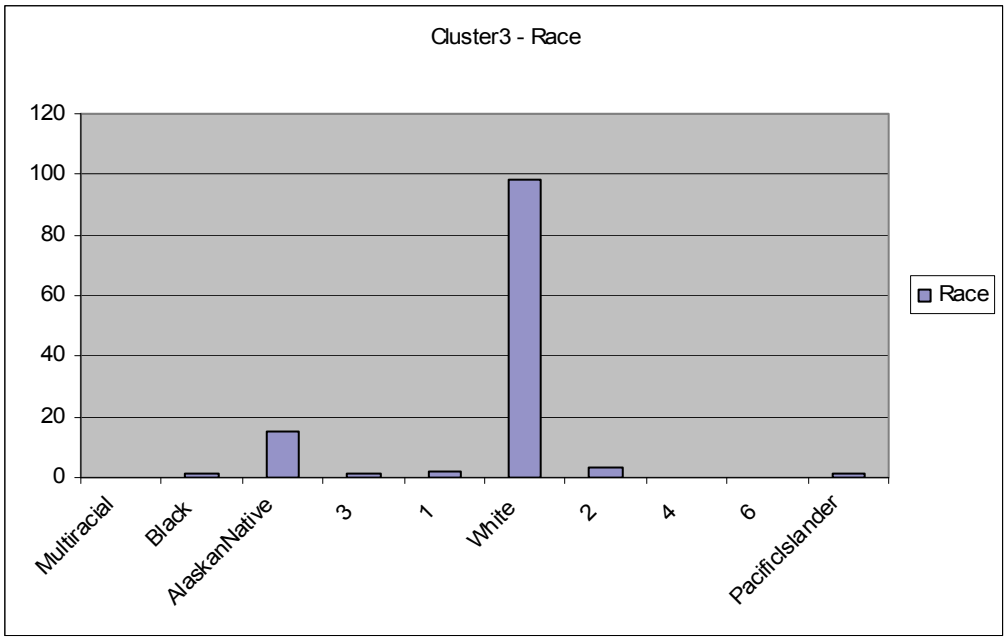

Figure B.15: K-Means Cluster 3 Race Distribution

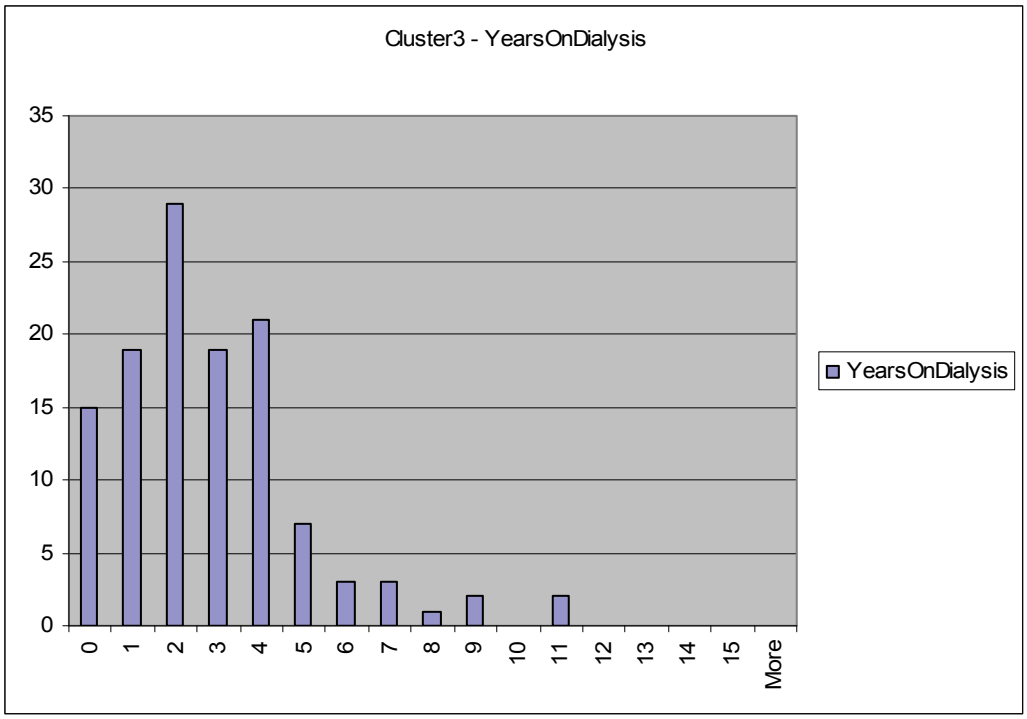

Figure B.16: K-Means Cluster 3 Years on Dialysis

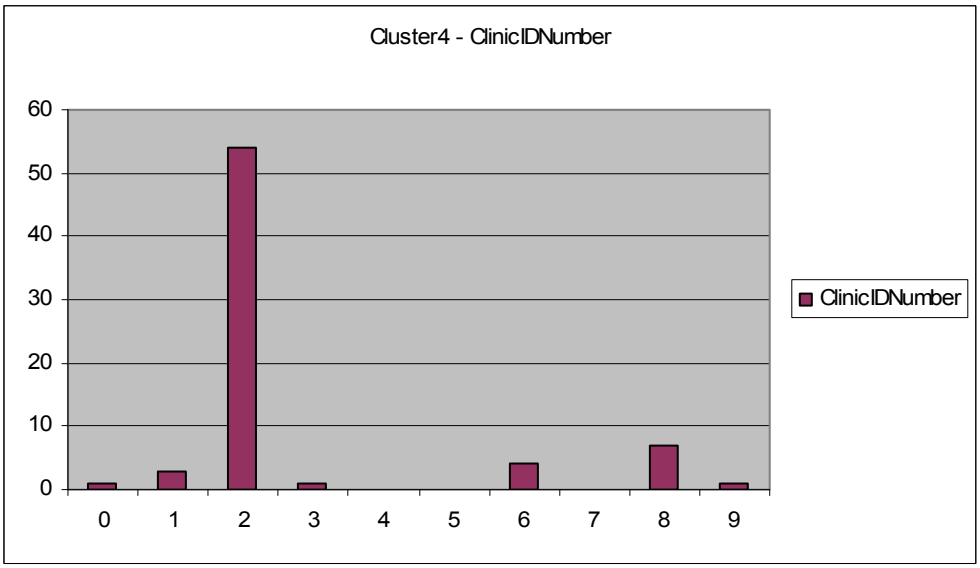

Figure B.17: K-Means Cluster 4 Clinic ID Number

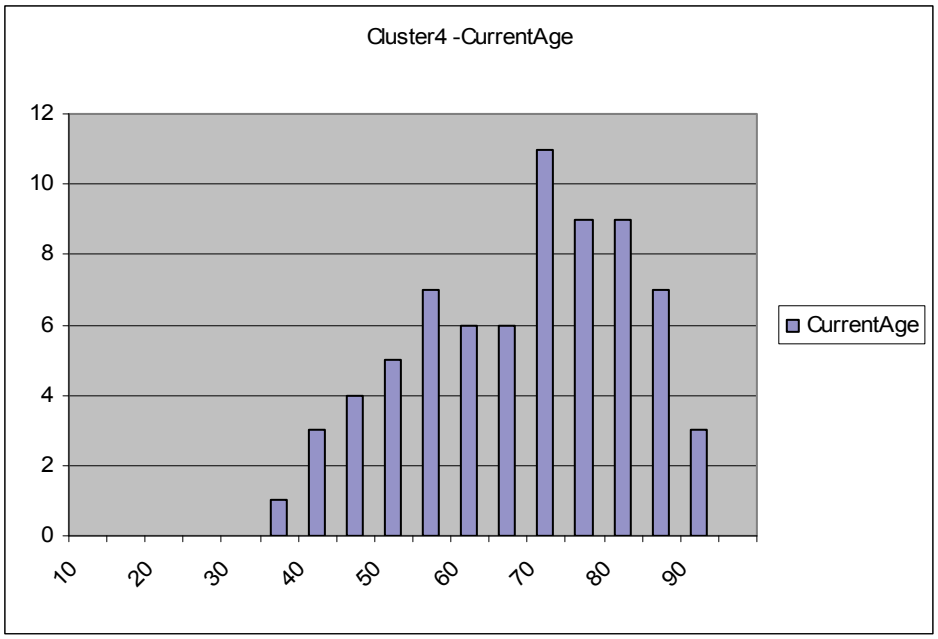

Figure B.18: K-Means Cluster 4 Age Distribution

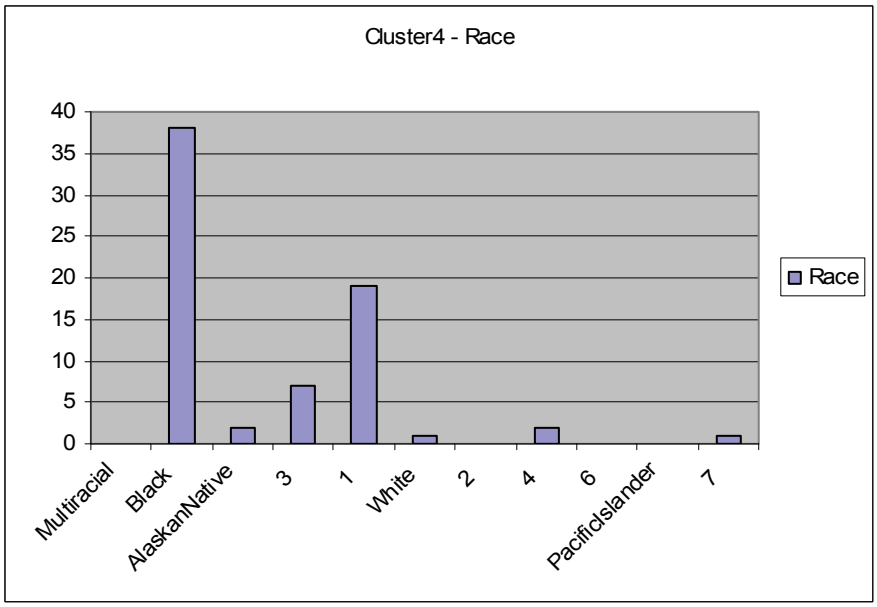

Figure B.19: K-Means Cluster 4 Race

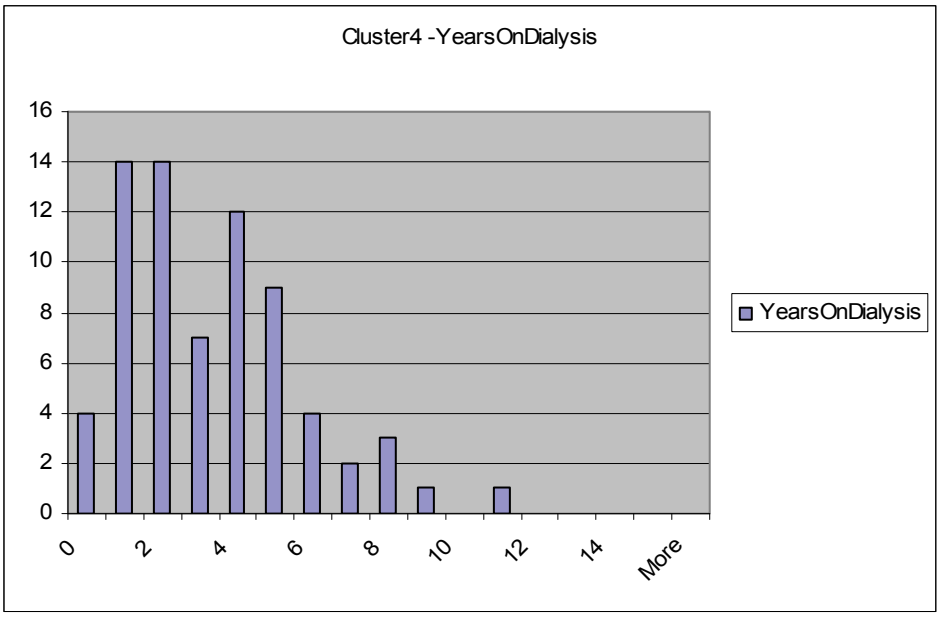

Figure B.20: K-Means Cluster 4 Years on Dialysis

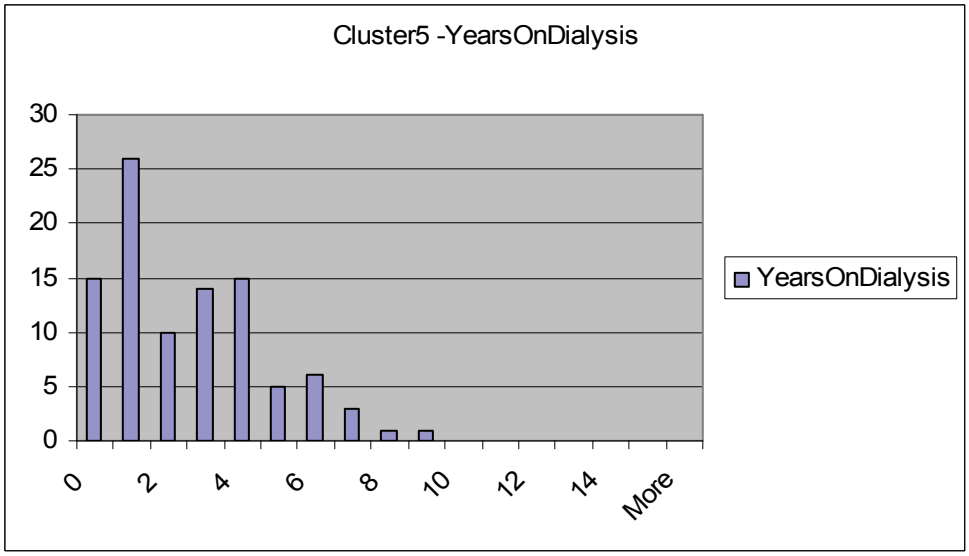

Figure B.21: K-Means Cluster 5 Years on Dialysis

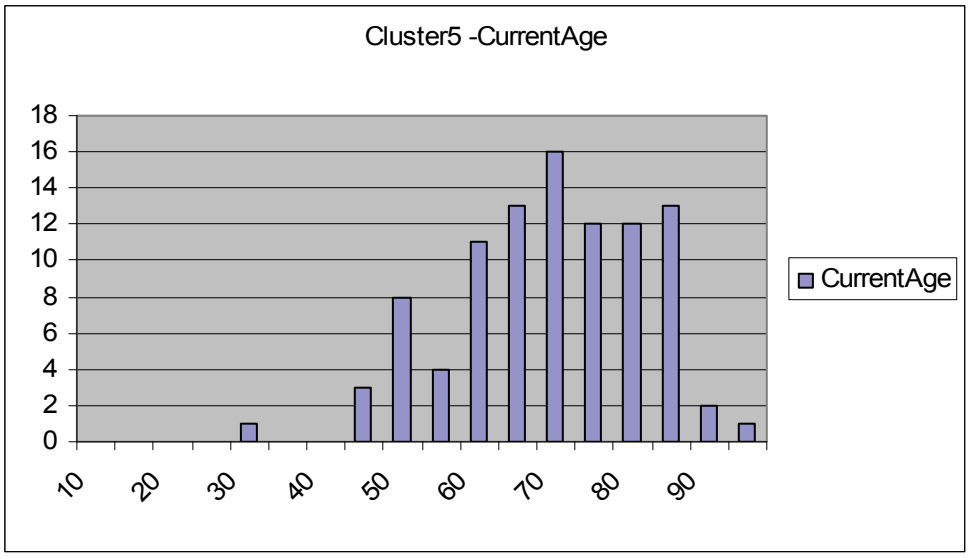

Figure B.22: K-Means Cluster 5 Age Distribution

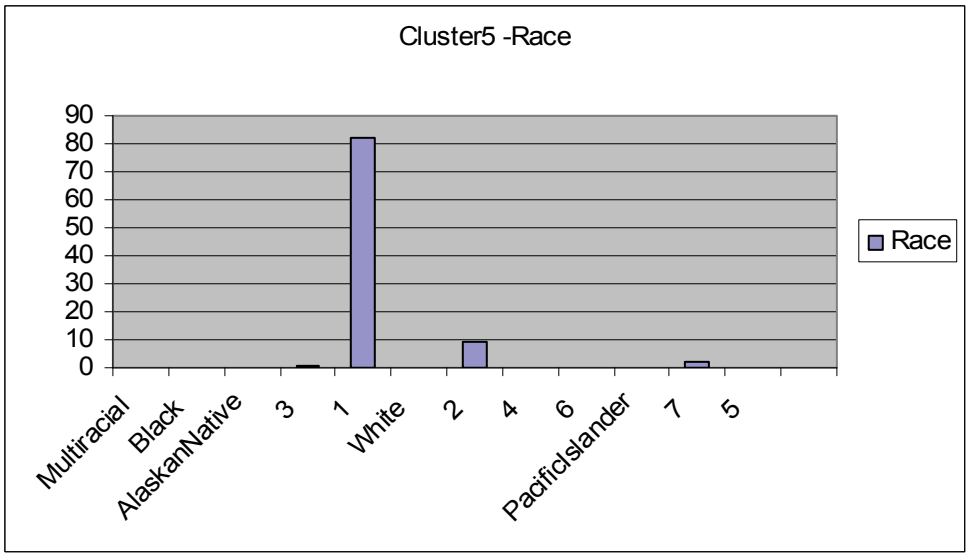

Figure B.23: K-Means Cluster 5 Race

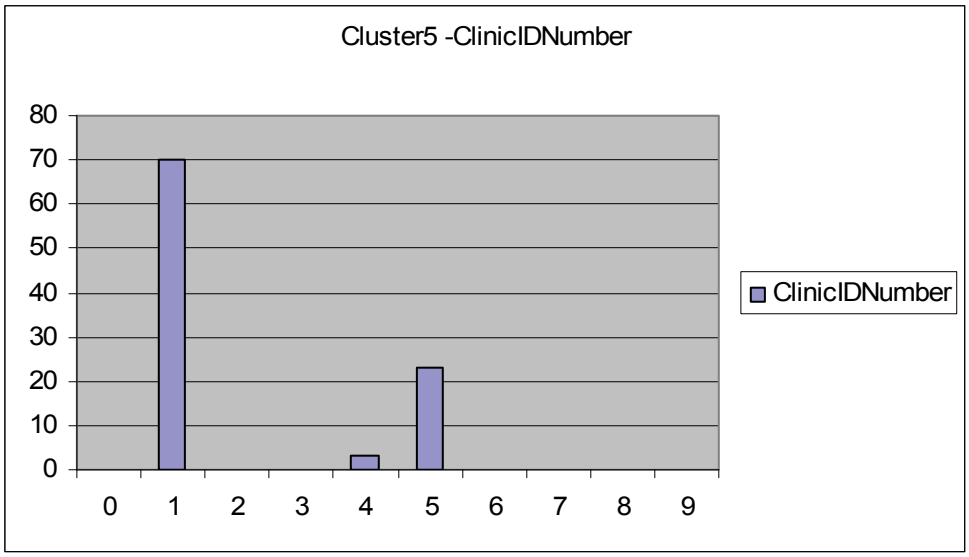

Figure B.24: K-Means Cluster 5 Clinic ID

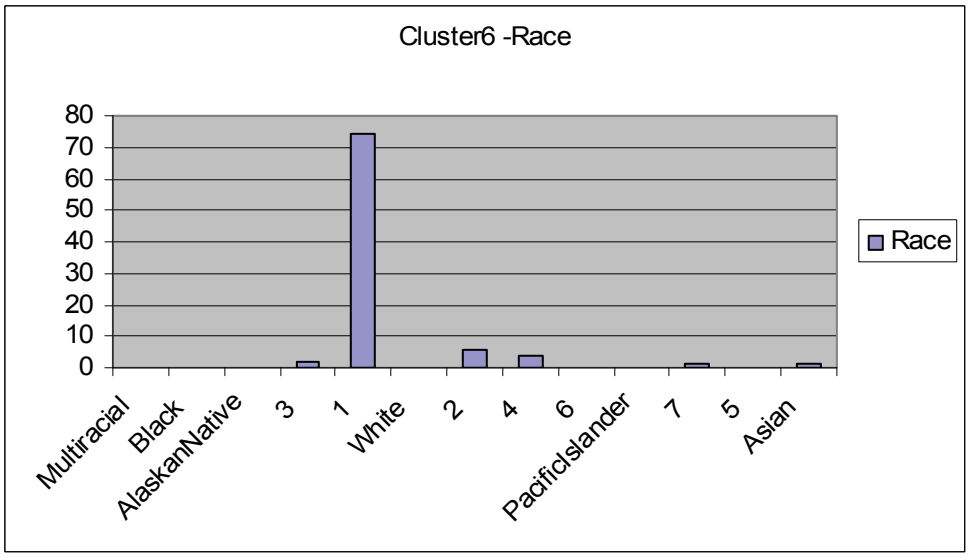

Figure B.25: K-Means Cluster 6 Race

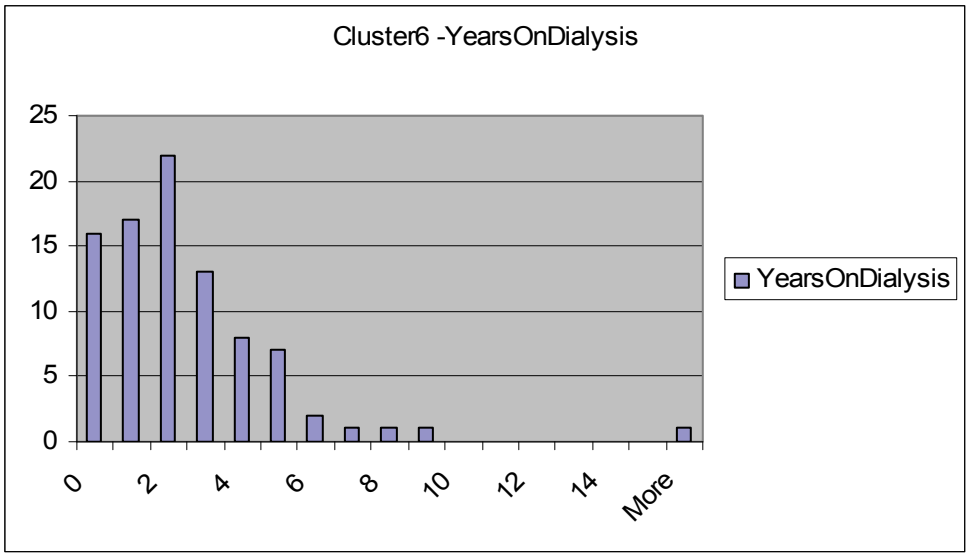

Figure B.26: K-Means Cluster 6 Years on Dialysis

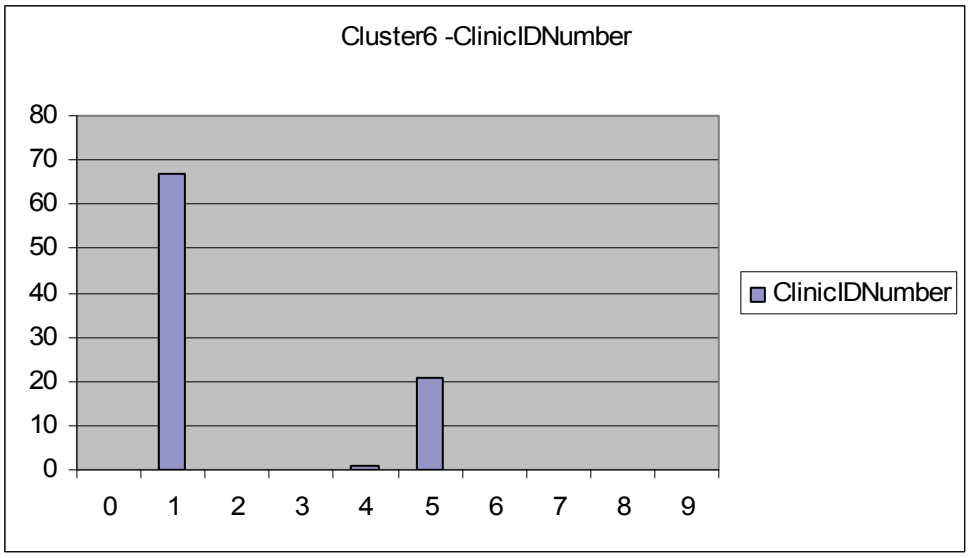

Figure B.27: K-Means Cluster 6 Clinic ID

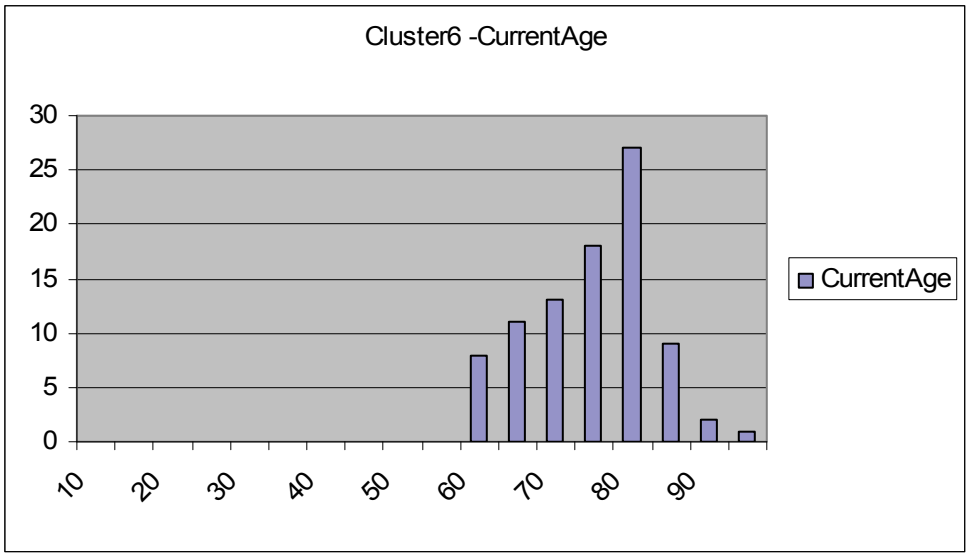

Figure B.28: K-Means Cluster 6 Age Distribution

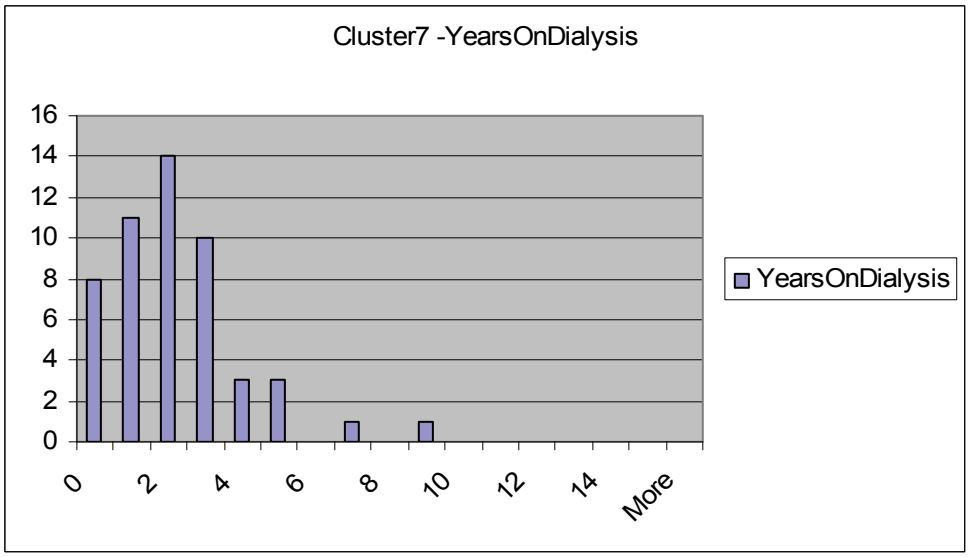

Figure B.29: K-Means Cluster 7 Years on Dialysis

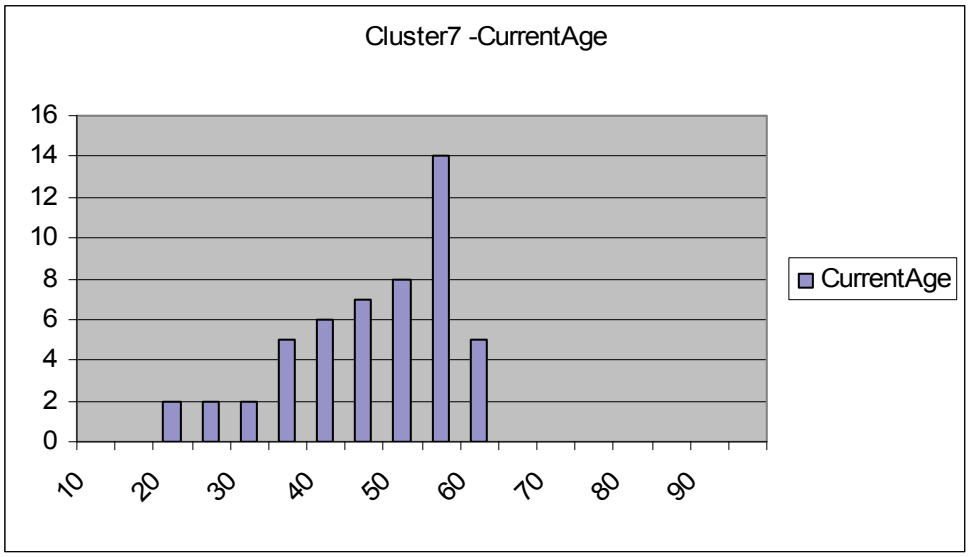

Figure B.30: K-Means Cluster 7 Age Distribution

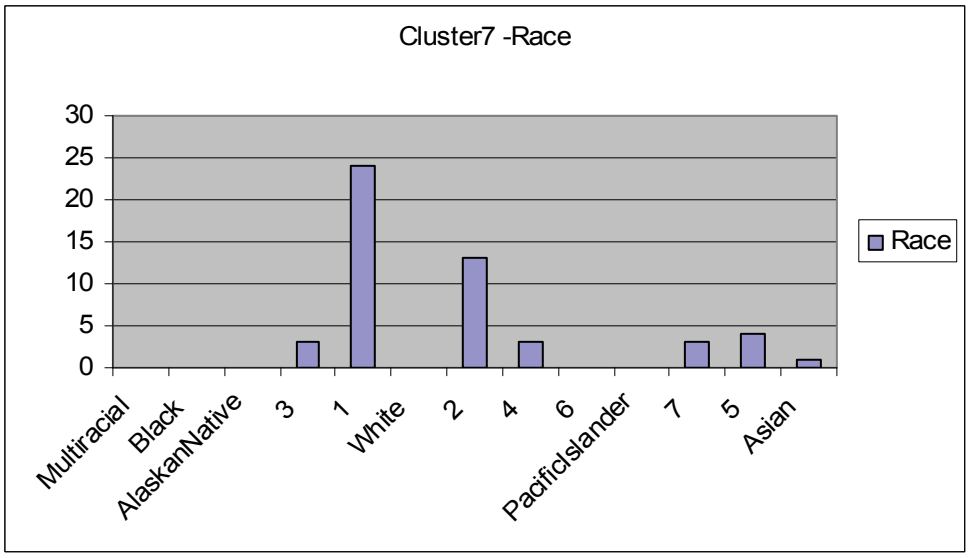

Figure B.31: K-Means Cluster 7 Race

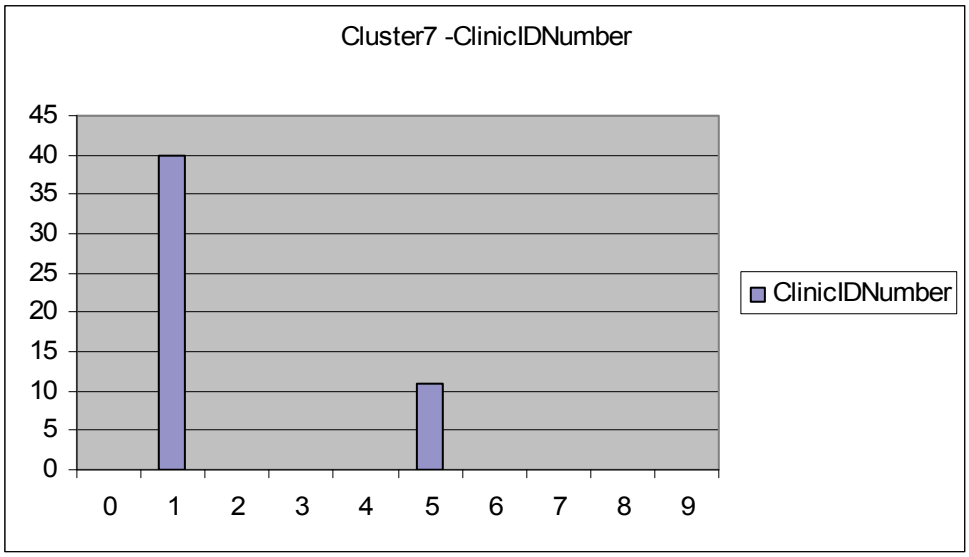

Figure B.32: K-Means Cluster 7 Clinic ID

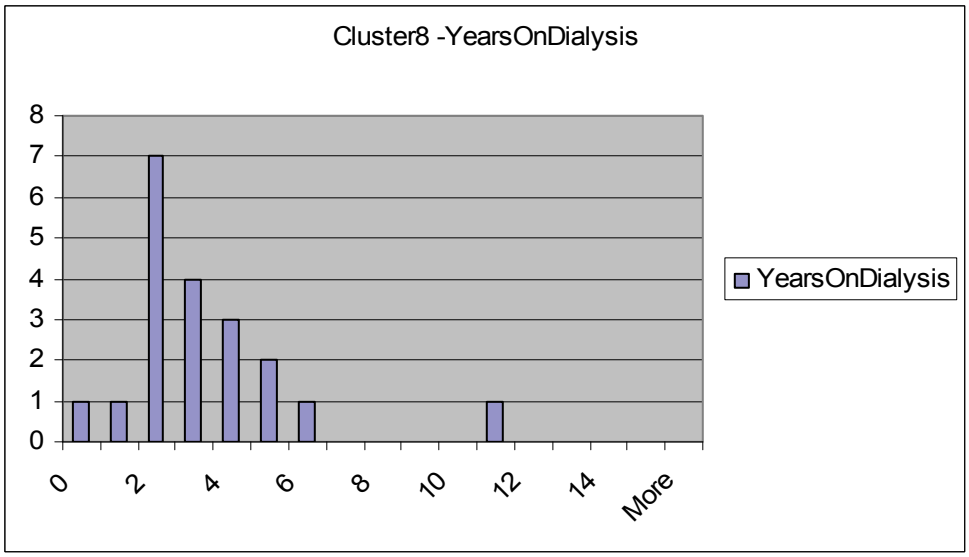

Figure B.33: K-Means Cluster 8 Years on Dialysis

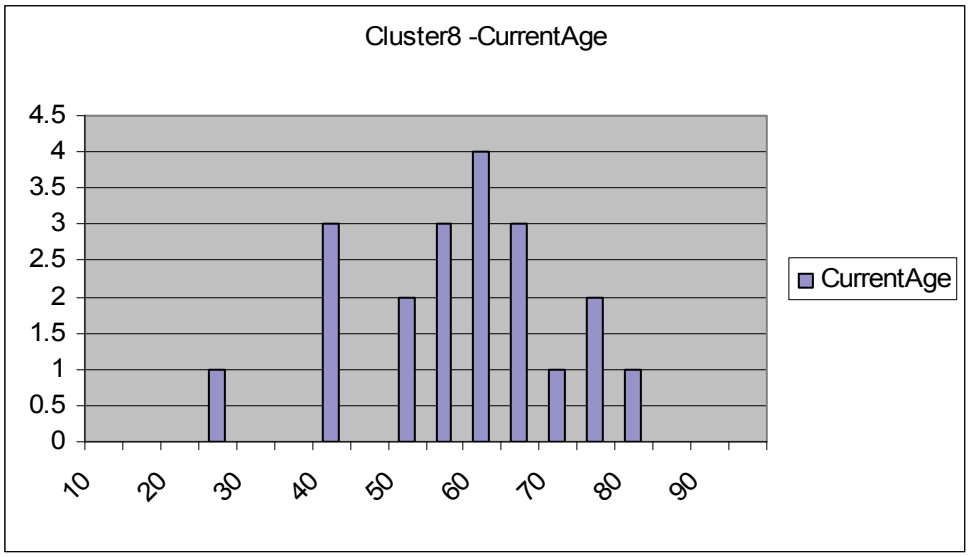

Figure B.34: K-Means Cluster 8 Age Distribution

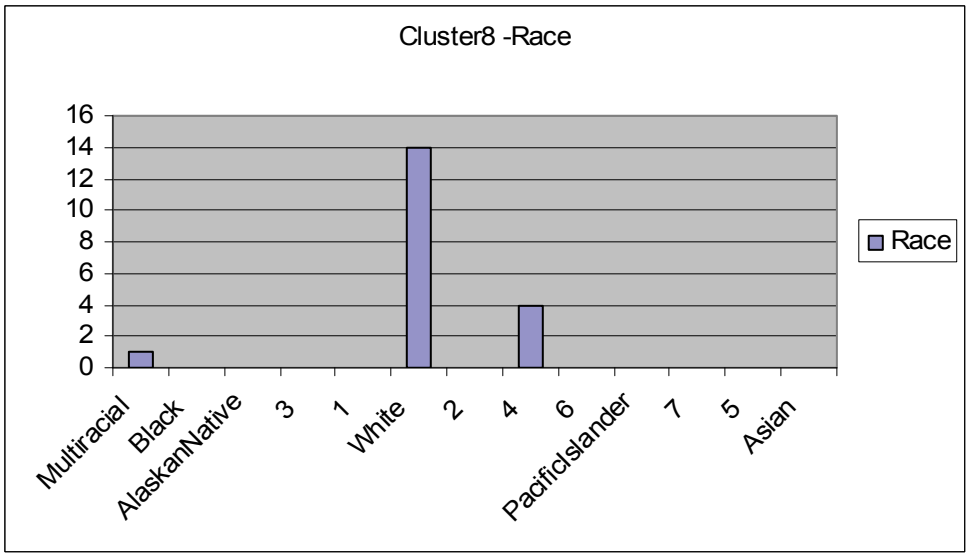

Figure B.35: K-Means Cluster 8 Race

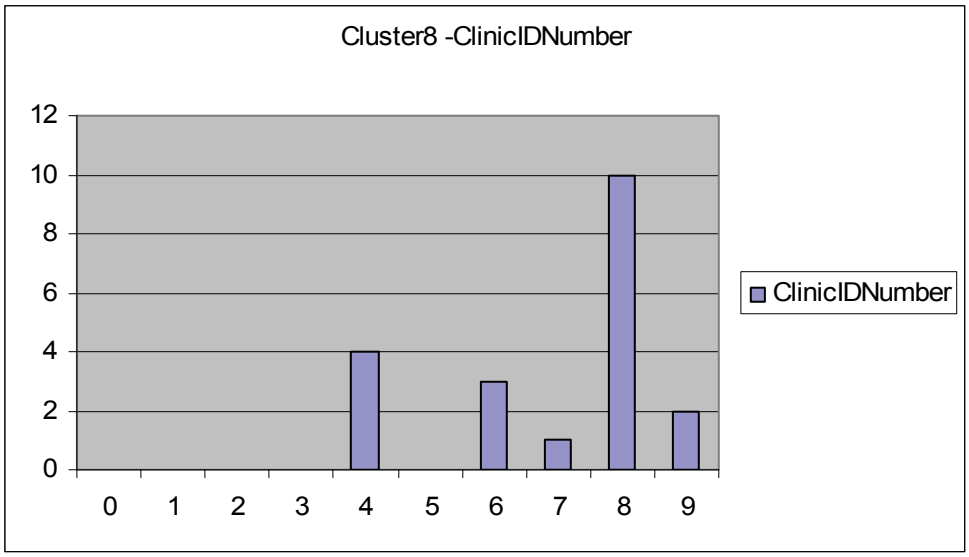

Figure B.36: K-Means Cluster 8 Clinic ID

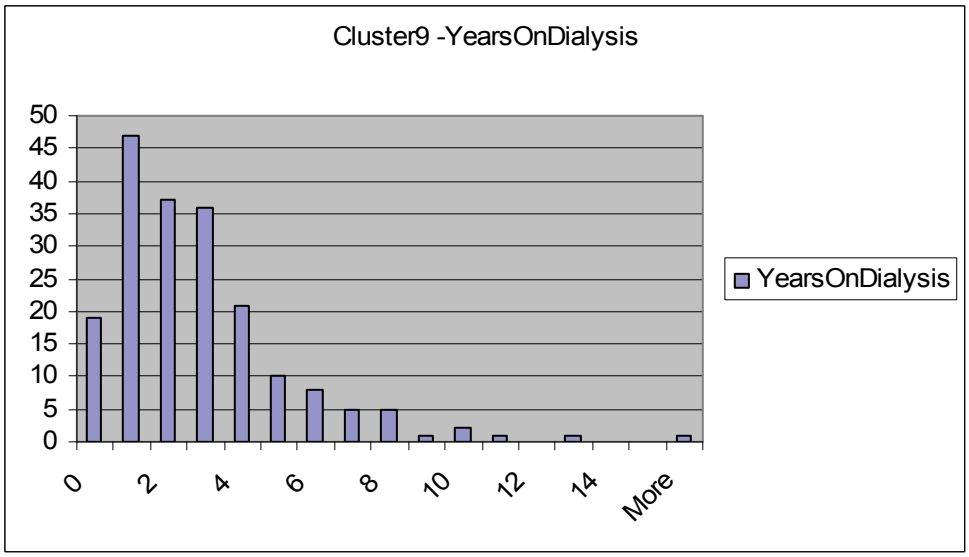

Figure B.37: K-Means Cluster 9 Years on Dialysis

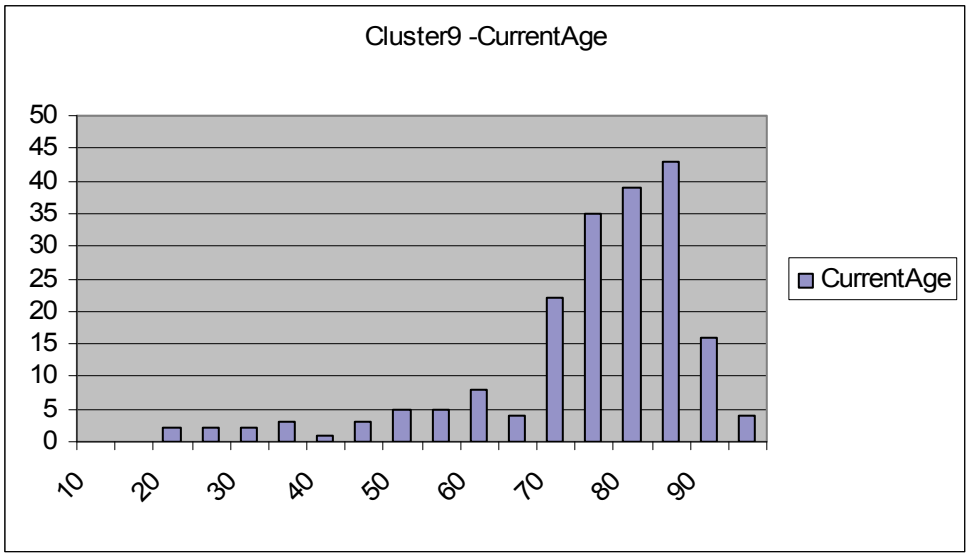

Figure B.38: K-Means Cluster 9 Age Distribution

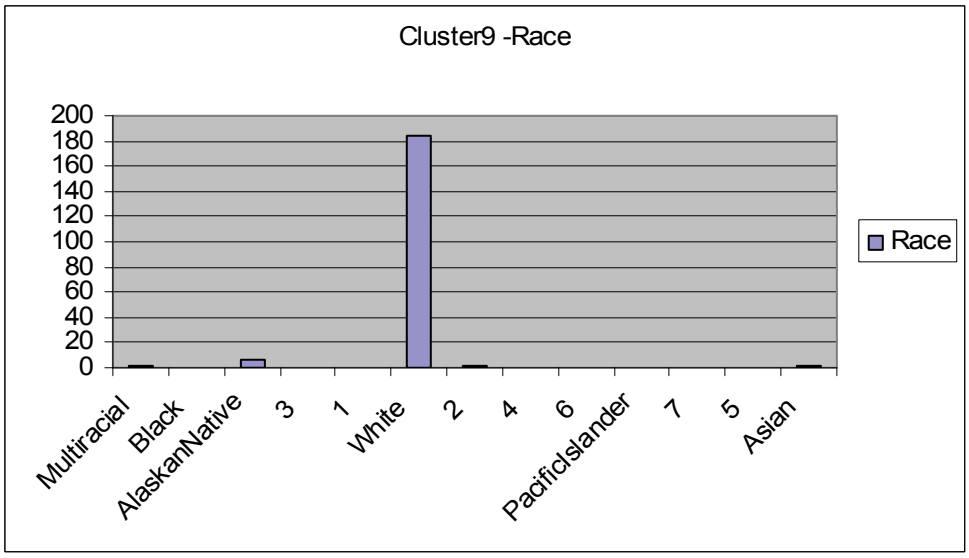

Figure B.39: K-Means Cluster 9 Race

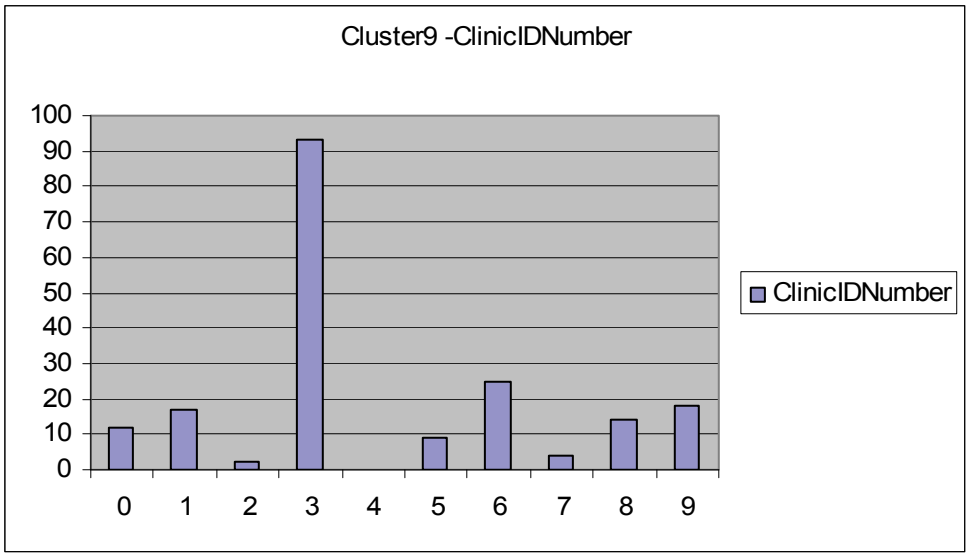

Figure B.40: K-Means Cluster 9 Clinic ID

APPENDIX C: EM CLUSTERING DISTRIBUTIONS

Table C.1: EM cluster assignment table (sample)

ClinicIDNumber	Ethnicity	Race	Gender	Age	YearsOnDialysis	ClusterAssignment
5	9	3	Male	18	0	cluster3
3	Non-Hispanic	White	Male	20	0	cluster0
3	Non-Hispanic	White	Male	20	2	cluster0
1	16	1	Male	20	1	cluster3
8	Hispanic-Mexican	White	Male	21	3	cluster1
0	Non-Hispanic	AlaskanNative	Female	21	2	cluster0
3	Non-Hispanic	White	Female	21	0	cluster0
6	Non-Hispanic	White	Male	21	1	cluster0
1	16	2	Male	21	2	cluster3
3	Non-Hispanic	Asian	Male	22	2	cluster0
2	Hispanic-Other	3	Female	23	2	cluster4
1	1	1	Male	23	4	cluster3
3	Non-Hispanic	White	Male	24	1	cluster0
4	Non-Hispanic	White	Male	24	0	cluster0
9	Non-Hispanic	White	Female	24	4	cluster1
4	Non-Hispanic	PacificIslander	Female	25	1	cluster4
4	Non-Hispanic	3	Male	27	3	cluster4
4	Non-Hispanic	White	Male	27	3	cluster0
1	Non-Hispanic	White	Male	27	1	cluster0

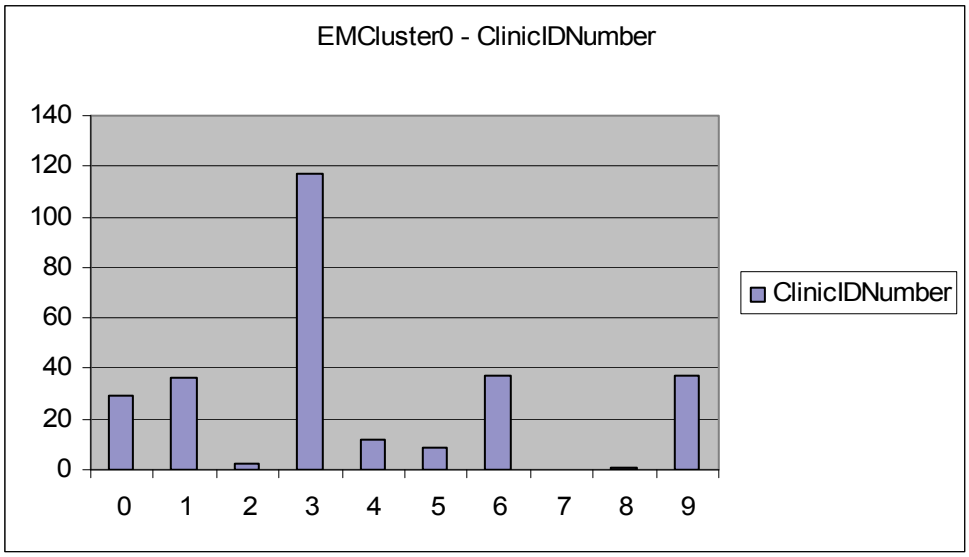

Figure C.1: EM cluster 0 Clinic ID

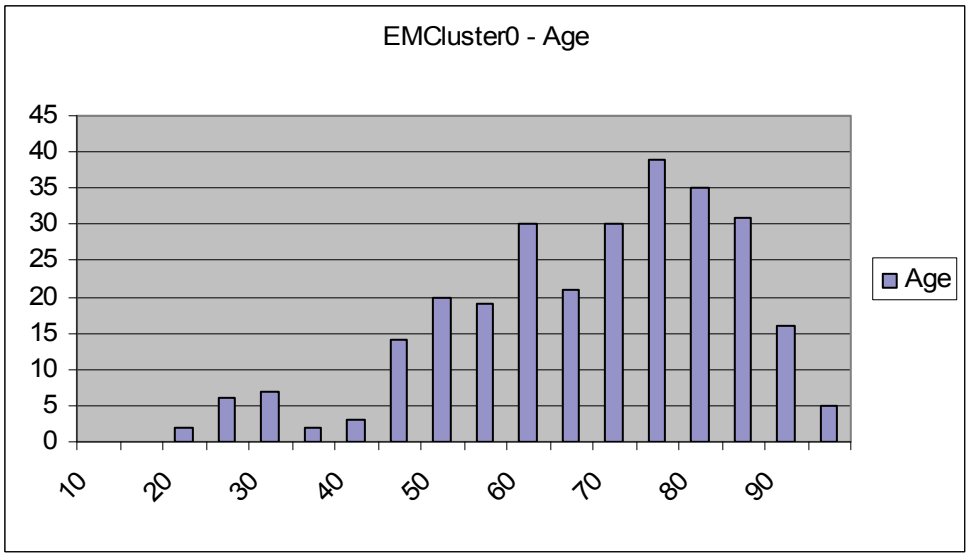

Figure C.2: EM cluster 0 Age Distribution

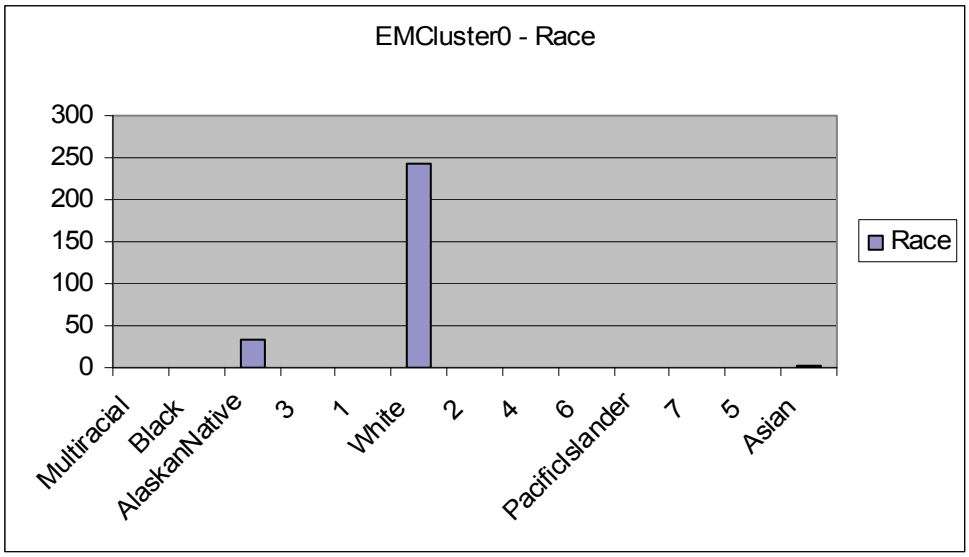

Figure C.3: EM cluster 0 Race

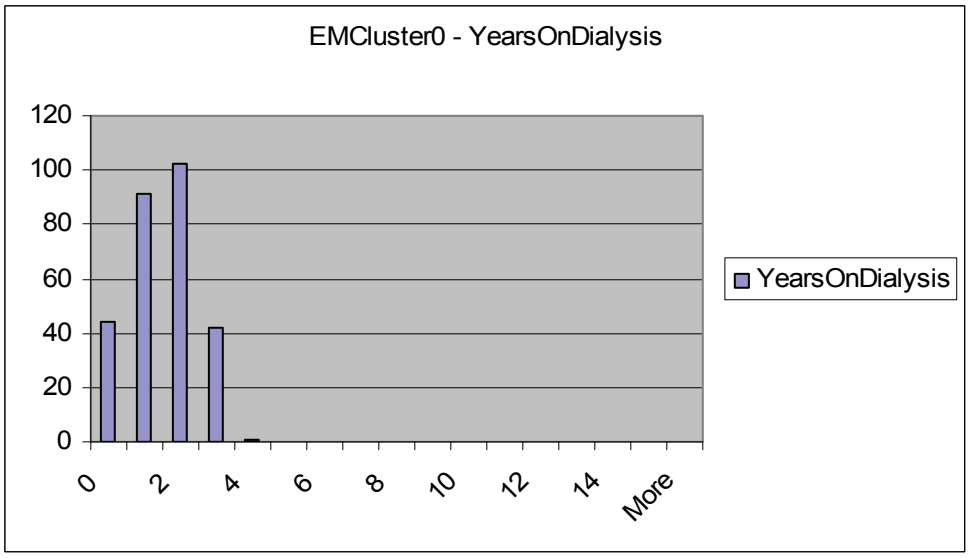

Figure C.4: EM cluster 0 Years on Dialysis

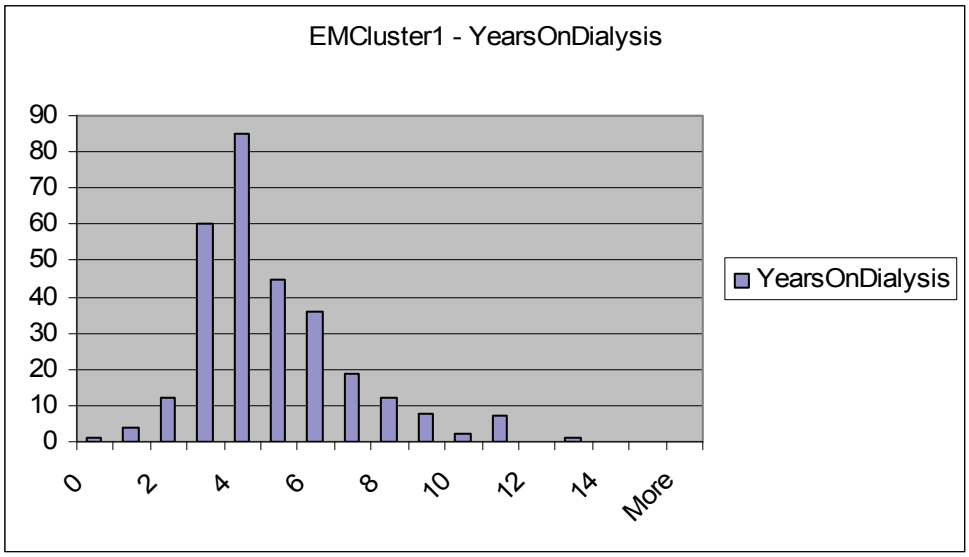

Figure C.5: EM cluster 1 Years on Dialysis

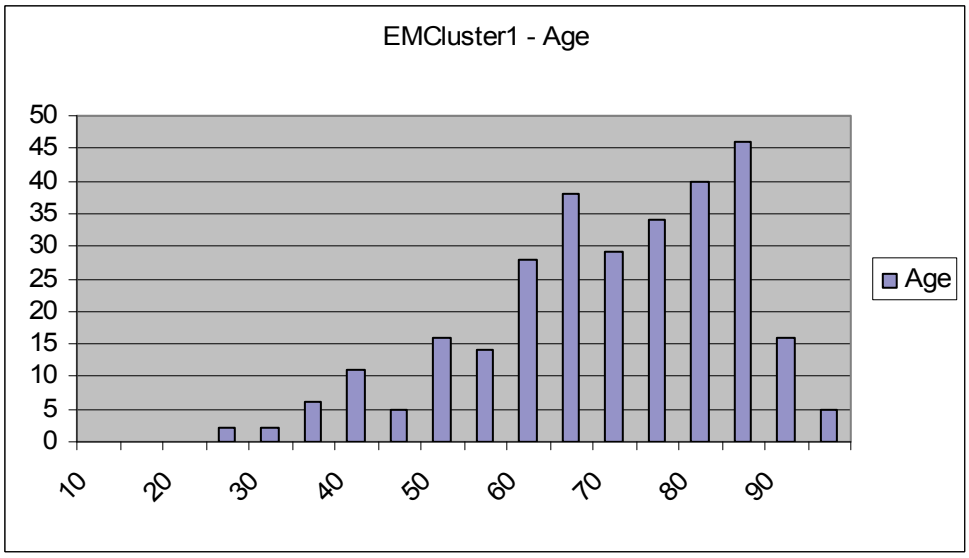

Figure C.6: EM cluster 1 Age Distribution

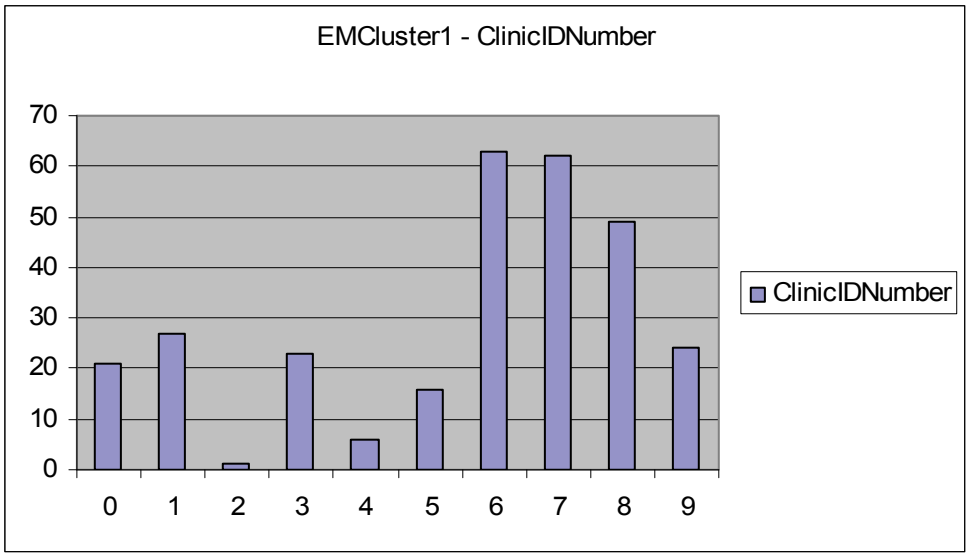

Figure C.7: EM cluster 1 Clinic ID

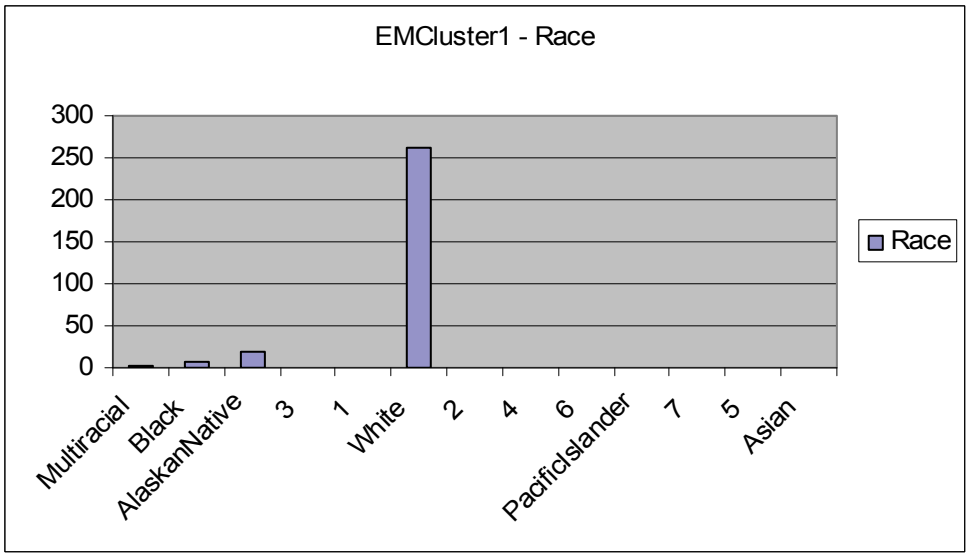

Figure C.8: EM cluster 1 Race

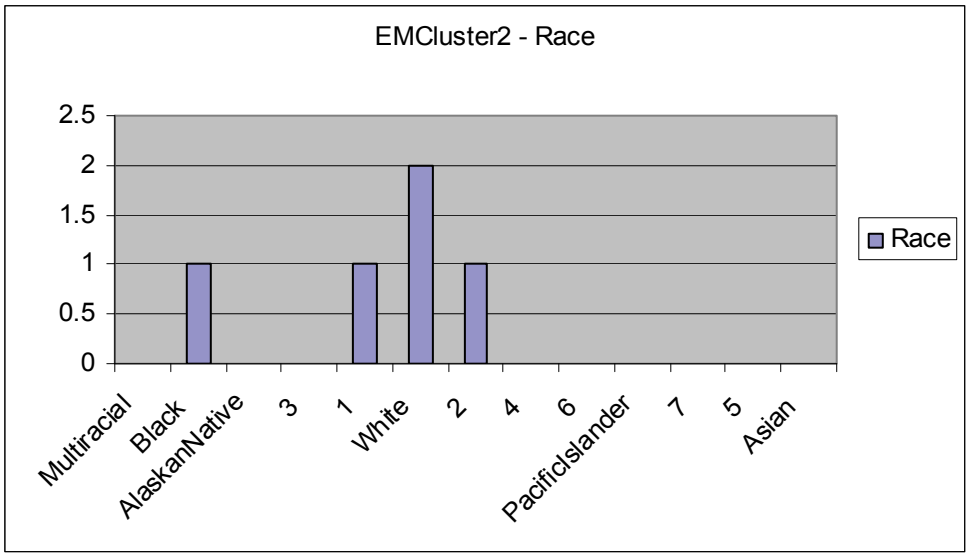

Figure C.9: EM Cluster 2 – Race

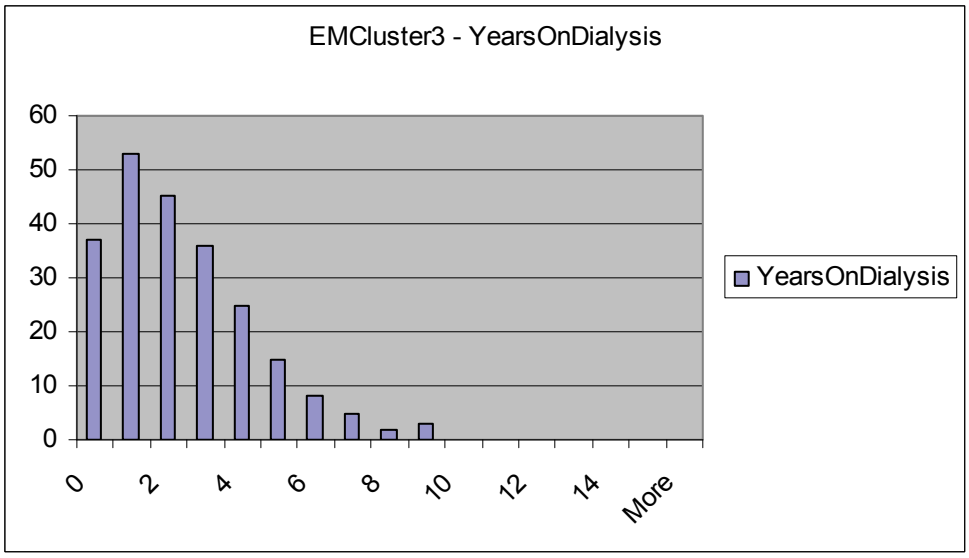

Figure C.10: EM Cluster 3 - Years on Dialysis

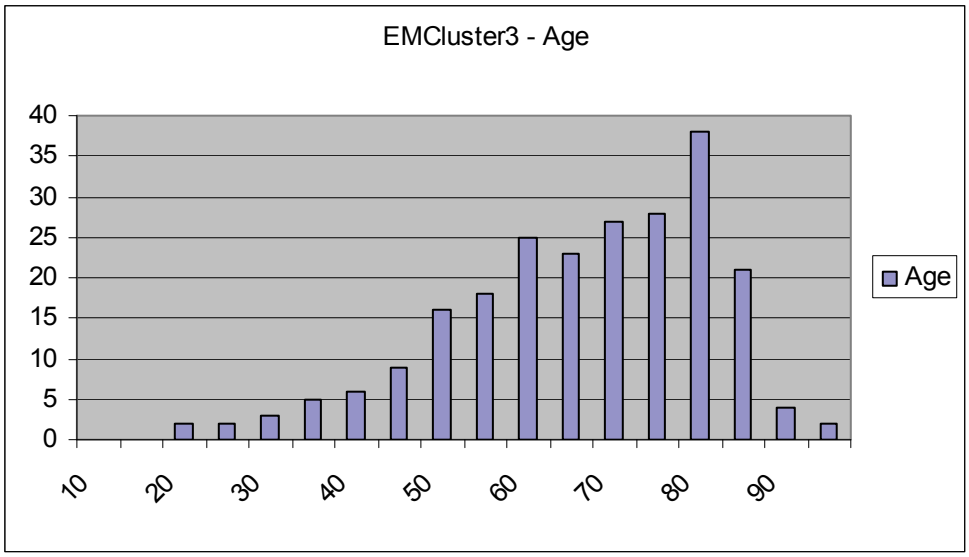

Figure C.11: EM Cluster 3 - Age Distribution

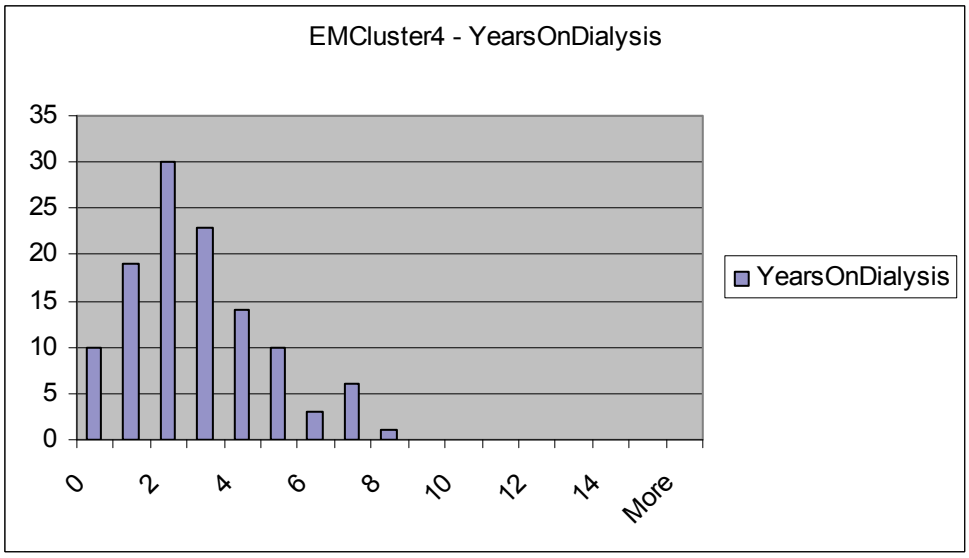

Figure C.12: EM Cluster 4 - Years on Dialysis

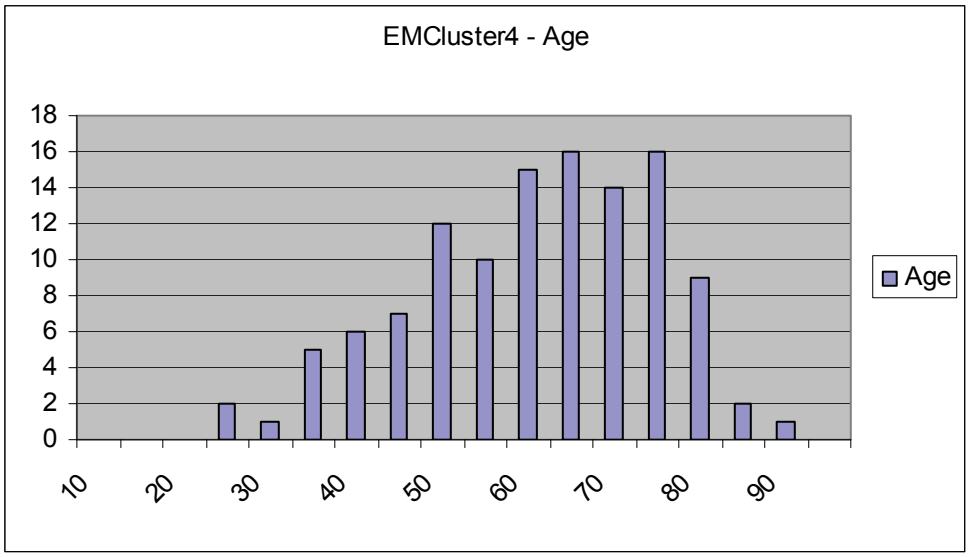

Figure C.13: EM Cluster 4 – Age

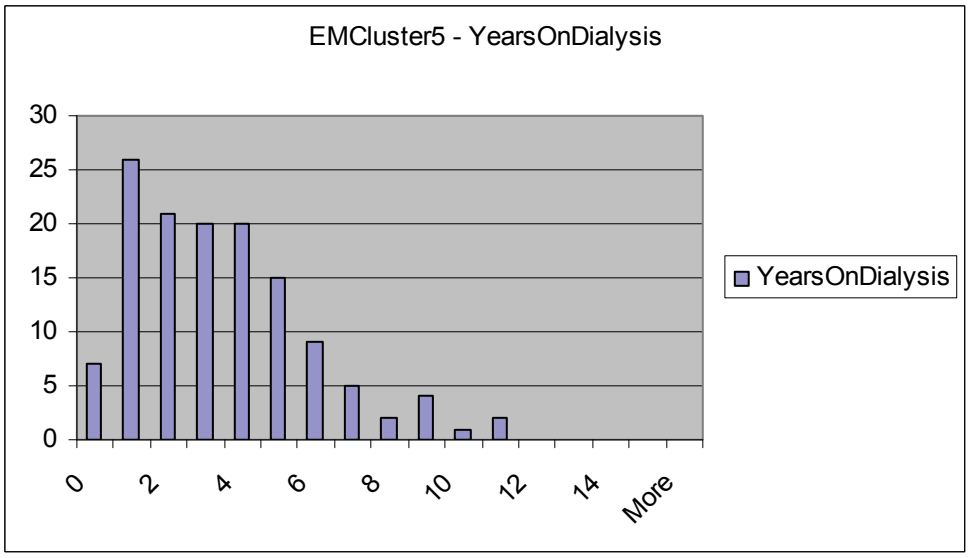

Figure C.14: EM Cluster 5 - Years on Dialysis

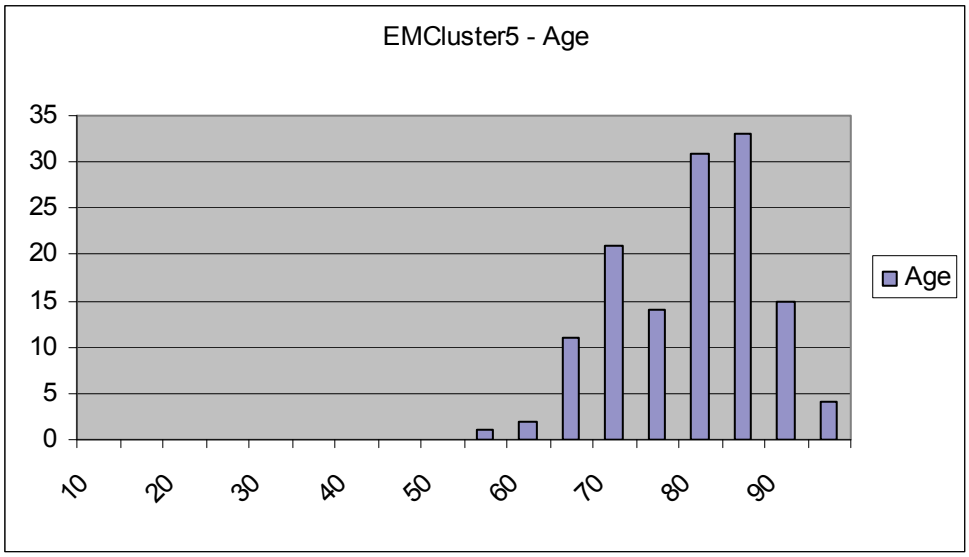

Figure C.15: EM Cluster 5 - Age

APPENDIX D: TIME-SERIES ANALYSIS

Table D.1: Full regressors set

Hgb-1	Hgb-2	Hgb-3	Hgb-4	Hgb-5	CurrEPO	EPO-1	EPO-2	EPO-3	EPO-4	EPO-5	CurrHgb
11.50	11.50	11.50	9.90	9.90	10000	12200	10800	12000	8400	4200	Normal
11.10	11.50	11.50	11.50	9.90	16000	10000	12200	10800	12000	8400	Low
10.40	11.10	11.50	11.50	11.50	18000	16000	10000	12200	10800	12000	Low
10.40	10.40	11.10	11.50	11.50	18000	18000	16000	10000	12200	10800	Low
8.80	10.40	10.40	11.10	11.50	24000	18000	18000	16000	10000	12200	Low
9.10	8.80	10.40	10.40	11.10	24000	24000	18000	18000	16000	10000	Low
9.10	9.10	8.80	10.40	10.40	24000	24000	24000	18000	18000	16000	Low
9.50	9.10	9.10	8.80	10.40	24000	24000	24000	24000	18000	18000	Low
9.50	9.50	9.10	9.10	8.80	26000	24000	24000	24000	24000	18000	Low
9.80	9.50	9.50	9.10	9.10	30000	26000	24000	24000	24000	24000	Low
9.80	9.80	9.50	9.50	9.10	30000	30000	26000	24000	24000	24000	Normal
11.00	9.80	9.80	9.50	9.50	30000	30000	30000	26000	24000	24000	Normal
11.00	11.00	9.80	9.80	9.50	30000	30000	30000	30000	26000	24000	Normal
11.10	11.00	11.00	9.80	9.80	30000	30000	30000	30000	30000	26000	Normal
11.10	11.10	11.00	11.00	9.80	30000	30000	30000	30000	30000	30000	Normal
11.70	11.10	11.10	11.00	11.00	30000	30000	30000	30000	30000	30000	Normal
11.70	11.70	11.10	11.10	11.00	30000	30000	30000	30000	30000	30000	Normal
11.70	11.70	11.70	11.10	11.10	28000	30000	30000	30000	30000	30000	High
12.50	11.70	11.70	11.70	11.10	24000	28000	30000	30000	30000	30000	High
12.50	12.50	11.70	11.70	11.70	24000	24000	28000	30000	30000	30000	High
12.50	12.50	12.50	11.70	11.70	24000	24000	24000	28000	30000	30000	High
12.50	12.50	12.50	12.50	11.70	22000	24000	24000	24000	28000	30000	High
12.50	12.50	12.50	12.50	12.50	18000	22000	24000	24000	24000	28000	High
12.50	12.50	12.50	12.50	12.50	18000	18000	22000	24000	24000	24000	High
12.30	12.50	12.50	12.50	12.50	18000	18000	18000	22000	24000	24000	High
12.30	12.30	12.50	12.50	12.50	16000	18000	18000	18000	22000	24000	High
12.90	12.30	12.30	12.50	12.50	12000	16000	18000	18000	18000	22000	High
12.90	12.90	12.30	12.30	12.50	13000	12000	16000	18000	18000	18000	Low
10.80	12.90	12.90	12.30	12.30	15000	13000	12000	16000	18000	18000	Low
9.80	10.80	12.90	12.90	12.30	24000	15000	13000	12000	16000	18000	Low
9.80	9.80	10.80	12.90	12.90	26000	24000	15000	13000	12000	16000	Low
9.70	9.80	9.80	10.80	12.90	30000	26000	24000	15000	13000	12000	Low
9.70	9.70	9.80	9.80	10.80	30000	30000	26000	24000	15000	13000	Low
9.90	9.70	9.70	9.80	9.80	30000	30000	30000	26000	24000	15000	Low