A Classification Model to Predict the Rate of Decline of Kidney Function

Ersoy Subasi[*]
Department of Engineering Systems
Florida Institute of Technology
150 W. University Blvd., Melbourne, FL 32901

Munevver Mine Subasi[†‡]
Department of Mathematical Sciences
Florida Institute of Technology
150 W. University Blvd.
Melbourne, FL 32901

Peter L. Hammer[§]
Rutgers Center for Operation Research
Rutgers University
640 Bartholomew Rd., Piscataway, NJ 08854

John Roboz[¶]
Department of Medicine
Icahn School of Medicine at Mount Sinai
One Gustave Levy Place, New York, NY 10029

Michael Lipkowitz[‖]
Department of Medicine
Georgetown University Medical Center
3800 Reservoir Rd., N.W., Washington D.C. 20007

Abstract

The African American Study of Chronic Kidney Disease and Hypertension (AASK), a randomized double-blinded treatment trial, was motivated by the high rate of hypertension-related renal disease in the African American population and the scarcity of effective therapies. This study describes a pattern based classification approach to predict the rate of decline of kidney function and to differentiate between proteomic samples of rapid and slow progressors. An accurate classification model consisting of 7 out of 5,751 serum proteomic features is constructed by applying the Logical Analysis of Data (LAD) methodology. The LAD discriminant is used to identify the patients in different risk groups. The LAD risk scores assigned to 116 AASK outperforms the risk scores assigned by proteinuria, one of the best predictors of chronic kidney disease.

1 Introduction

Chronic kidney disease (CKD), defined by reduced glomerular filtration rate (GFR), proteinuria, or structural kidney disease, is a worldwide growing public health problem[1]. Many subjects with renal disease of most etiologies progress to renal failure and end stage renal disease (ESRD), requiring renal replacement therapy, which may involve a form of dialysis or renal transplantation (Lewis et al. 1993; Klahr et al. 1994; DCCT 1995; Brenner et al. 2001; Lewis et al. 2001; Wright et al. 2002; Niki, Panos, and Christos 2015). Identification and characterization of novel biomarkers and targets of therapy for the ESRD patients remains the focus of the current research in the field of critical care medicine and has been the objective of a number of studies such as the Chronic Renal Insufficiency Cohort (CRIC)[2] established in 2001 by the National Institute of Diabetes, Digestive, and Kidney Diseases (NIDDK) to improve the understanding of CKD and related cardiovascular illness and the African American Study of Kidney Disease and Hypertension (AASK) Cohort which examines traditional and non-traditional risk factors for progression of CKD. AASK, a randomized double-blinded treatment trial, was motivated by the high rate of hypertension-related renal disease in the African American population and the scarcity of effective therapies. The study began as a 21-center randomized double-blinded treatment trial of 1,094 African Americans patients, aged 18-70 yrs with hypertension and renal failure with GFR between 11.6 and 37.6 $mL/min/m^2$ (average body surface: 1.73 m^2). Other known causes of renal disease such as diabetes were exclusion criteria as was proteinuria $> 2.5gm/gm$ creatinine. The initial AASK results were mixed (Wright et al. 2002). It was shown that while therapy can slow the progression of renal disease, there was still high rate of progression to renal failure. The calcium channel blocker (dihydropyridine) arm of the study was stopped early when interim analysis indicated that it was inferior to both the beta-blocker (metoprolol) and angiotensin converting enzyme treatment (ramipril) arms (Agodoa et al. 2001). After the clinical trial phase, AASK re-enrolled 691 of the surviving into the observational AASK Cohort study. Follow-up for 10 years on these subjects showed that despite best practice treatment with ramipril and a BP goal of 130/80 more than half the patients had a doubling of creatinine, ESRD or death (Appel et al. 2008)). Several possible interventions, including controlling blood pressure (Wright et al. 2002), treating diabetes (DCCT 1995), modifying dietary protein intake (Klahr et al. 1994) and using medications that might have renoprotective effects (Wright et al. 2002; Agodoa et al. 2001; Ruggenenti et al. 1999) have been tested in clinical trials. In all cases, the residual rate of progression of renal disease has remained significant. To date there are few prediction models to identify which patients are likely to progress significantly. The present paper describes a proteomics approach, using a mass spectrometric technique, attempting to find putative prognostic markers for disease progression. The technique used was surface-enhanced laser desorption ionization/time of flight (SELDI-TOF) mass spectrometry to obtain a spectrum of serum proteins with masses from several hundred to about $20kDa$. We

[*] Email: esubasi@fit.edu

[†] Email: msubasi@fit.edu

[‡] Corresponding author.

[§] Deceased.

[¶] Email:john.roboz@mssm.edu

[‖] Email: Michael.S.Lipkowitz@gunet.georgetown.edu

[1] Chronic Kidney Disease Surveillance Project, Center for Disease Control and Prevention – http://nccd.cdc.gov/ckd/

[2] http://www.cristudy.org/Chronic-Kidney-Disease/Chronic-Renal-Insufficiency-Cohort-Study/

use a powerful and robust classification method, called Logical Analysis of Data, to obtain serum proteomic patterns that can accurately identify patients at high risk of progression of CKD. The ultimate goal of the future studies motivated by the results of this paper is to identify the proteins to define new targets for interventions.

2 Study Subjects and Methods

There was significant heterogeneity of progression rate of renal disease in the AASK Trial as can be seen in Figure 1. The rate of decline of GFR after 6 months in the trial (chronic GFR slope) is depicted in blue for each patient from most rapid decline (negative slope) on the left, to the least rapid (positive slope) on the right of the Figure. It is generally assumed that the expected rate of decline of GFR with aging is $-1ml/min/yr$ (Berg 2006; Murussi, Gross, and Silveiro 2006), although longitudinal studies have raised questions about this assumption (Lindeman, Tobin, and Shock 1985; Lindeman 1990). Using this estimate, approximately 30% of the patients in Figure 1 did not progress (right side, slope $> -1ml/min/yr$) while approximately 30% progressed rapidly (left side, slope $< -3ml/min/yr$). Of interest, it is also apparent that proteinuria, while the strongest predictor of progression rate in most studies, is not an ideal predictor in that there are a number of slow progressors with significant proteinuria (red spikes, right), while a significant number of rapid progressors had no or minimal proteinuria (absence of red bars, left). This data is supported by the observation in genetics studies that proteinuria and progression of renal disease may be disparate phenotypes (Fogarty et al. 2000; Krolewski et al. 2006).

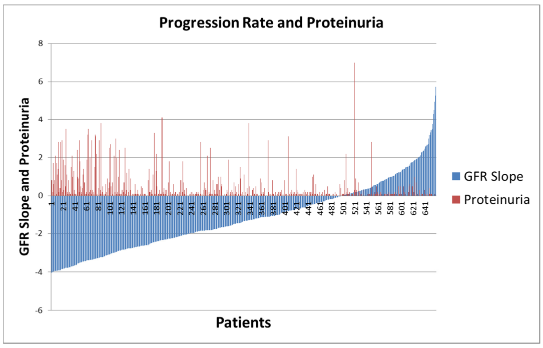

Figure 1: Patients stratified by GFR slope (blue bars) with degree of proteinuria superimposed (red spikes)

2.1 Sample Preparation

The sera samples from the study subjects were aliquoted into 5-8 L sizes so that sera were not thawed more than twice. After adding internal standard (insulin, 5734.5 Da), binding buffer (PBS, containing 25 mM imidazole, pH 7), mixing and centrifuging, 5 L aliquots were placed on metal chelation ProteinChips (IMAC30) charged with copper and prepared as suggested by the manufacturer (Cyphergen Biosystems, Fremont CA) using 100 mM CuSO4 for 15 min. The energy-absorbing matrix was -cyano-4-hydroxy-cinnamic acid (sinapinic acid). Samples were analyzed in a random fashion, with both rapid and slow progressors assayed on each 8-spot chip. Samples were analyzed in duplicate. The operators were blinded to the identity of the samples.

2.2 Analysis using SELDI-TOF-MS

Masses (range 0-20,000 Da) and intensities were determined using a Protein Biology System 2 SELDI-TOF mass spectrometer (Ciphergen Biosystems, Freemont, CA) and Protein Chip Biomarker software (version 3.0, Ciphergen Biosystems). The laser intensity was adjusted depending on the signal to noise ratio of the mass peak heights to background. The spectra were generated at using signal averaging 90 laser shots. Size and amplitude controls were run daily. A series of proteins with known masses were used as external mass calibrators. Peak intensities were normalized to the internal standard, insulin, 5734.5 Da.

2.3 Study Subjects

We have performed a pilot study on a selected subset of subjects from the African American Study of Kidney Disease and Hypertension (AASK) Clinical Trial based on the glomerural filtration slope (GFR) of all AASK patients presented in Figure 1. Two sets of subjects were selected from the AASK study: “rapid progressors” (most rapid quartile of GFR slope after 3 months, 57-patients) and “slow progressors” (slowest quartile of GFR slope after 3 months, 59-patients) based on the GFR histogram presented in Figure 2.

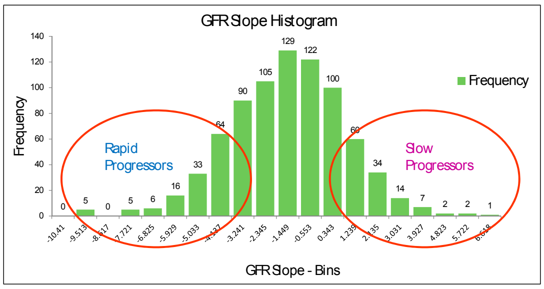

Figure 2: GFR Slope of AASK Patients

Aliquot of serum from the 1-year visit for each subject were processed in a standardized protocol, used for clinical assays, and then frozen at -80 C. Patients were matched for AASK blood pressure goal and medication. All subjects and research related protocols were evaluated and approve by the IRB of Mount Sinai School of Medicine. Signed Informed Consent was obtained before the start of the AASK study in all participating subjects.

Table 1 describes the patient population for this pilot study. Fast progressors were more likely male, had lower measured glomerular filtration rate, and a higher degree of proteinuria. GFR slope was disparate as expected. While GFR was significantly different, the magnitude of the difference was relatively small and would not be expected to dramatically affect the serum proteome.

	Slow (n=59)	Rapid (n=57)	p-value
GFR Slope	2.18 ± 1.13	-6.64 ± 1.38	< 0.0001
GFR	53.6 ± 11.6	45.3 ± 12.2	< 0.001
Proteinuria	0.10 ± 0.19	1.09 ± 1.37	< 0.0001
Age	53.4 ± 11.6	51.2 ± 12.1	NS
BMI	30.0 ± 6.0	32.0 ± 7.4	NS
Male	29	39	< 0.001
Female	30	18	

Table 1: Data Characteristics

Proteinuria was also very different between the 2 groups. This is consistent with prior data (Wang et al. 2006) that showed that proteinuria is the strongest predictor of GFR slope progression in AASK. The discrepancy in gender was fortuitous and unexpected, since gender did not predict GFR slope in AASK (Wang et al. 2006).

The SELDI data were accurate and reproducible. The detected mass for insulin in the 116 samples was $5737.3 + 1.1$, which was within 0.06% of the MW of insulin, which is within expected ranges for SELDI. The mean intensity of the insulin peak was $28.96 + 2.02$. All intensities were subsequently normalized using the internal insulin standard before any applying data analysis techniques. The results for each patient were highly reproducible, with the mean correlation coefficient between runs for each patient $0.820 + 0.115$. The mean of the SELDI-TOF spectra for 57 rapid and 59 slow progressors is depicted in Figure 3. We eliminated data with $m/z < 500$ as we do not believe peptides with masses this small will be meaningful, and data with $m/z > 12,000$ as the intensities above this value dropped significantly, and accuracy of the SELDI reader at larger masses is limited. This resulted in a dataset with 116 AASK patients (57 rapid and 59 slow progressors) with 5,751 serum proteomic features.

We observed that a number of peaks differed between rapid and slow progressors. The mean data for rapid and slow progresors was highly reproducible (correlation coefficient > 0.98), and of interest, most of the peaks that differed were greater in the slow progressors, suggesting that rapid progression may relate to loss/absence of protective proteins/peptides.

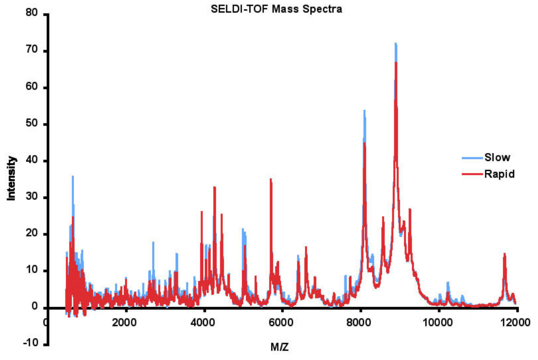

Figure 3: SELDI-TOF mass sprectra of slow (blue) vs rapid (red) progressors

2.4 Logical Analysis of Data

Extracting meaningful information from a large-scale data such as AASK presents requires the integration of sophisticated mathematical techniques and efficient computerized algorithms. Several studies have applied existing algorithms for pattern matching to SELDI datasets (Li et al. 2002; Qu et al. 2002; Ball et al. 2002), but no significant attempts to optimize such methods have been attempted. In this study we applied the Logical Analysis of Data (LAD) methodology to build a classification model based on the serum proteomic feature for distinguishing the rapid progressors from the slow progressors in AASK dataset. LAD is a two-class classification method based on the theory of Boolean functions, optimization, and combinatorics. LAD was shown to outperform previously employed methods on publicly available datasets (Alexe et al. 2004; 2005; 2006a; Reddy et al. 2008).

First introduced in (Crama, Ibaraki, and Hammer 1988; Boros et al. 2000), LAD has been successfully applied in various areas of science and technology. The medical applications of LAD (Alexe et al. 2003; Lauer et al. 2002; Alexe et al. 2004; 2005; 2006a; 2006b; Hammer and Bonates 2006; Reddy et al. 2008) demonstrate clearly that the method provides excellent solutions both for the analysis of medical problems using clinical datasets, and for that of multiparameter datasets generated in the fields of genomics (such as gene expression microarrays (Alexe et al. 2006b)) and proteomics (such as mass spectrometry (Alexe et al. 2004; 2005; 2006a; Reddy et al. 2008)). The results of LAD applied to such problems turned out not only to be extremely accurate, but also to yield new kinds of tools both for direct applications, such as diagnosis and prognosis, and for biomedical research about the importance of targeting particular combinations of genes and proteins for therapeutic purposes.

LAD is a multistep procedure consisting of (1) discretization, (2) support set selection, (3) pattern generation, (4) classification, and (5) cross-validation. The goal of LAD as applied in this case to rate of progression of chronic kidney disease is to detect patterns, or “combinatorial biomarkers”, consisting of restrictions imposed on the values of the intensities of a combination of several masses in the SELDI data. The technique generates patterns nearly exhaustively and in an algorithmically efficient way, and then uses a collection of patterns in a “classification model” that can predict progression rate of the AASK patients selected in this study. In contrast to many statistically based methods, the procedures of LAD are guided by the collective strength of the subsets of features (protein/peptide peaks), without being necessarily restricted to those which have the highest individual correlation coefficients with the outcome.

The initial step of LAD is the transformation of numeric features into binary features without losing predictive power. The procedure consists of finding cut-points for each numeric feature. The set of cut-points can be interpreted as a sequence of threshold values collectively used to build a global classification model over all features (Boros et al. 2000; Hammer and Bonates 2006). Discretization is a very useful step in data-mining, especially for the analysis of medical data (which is very noisy and includes measurement errors) – it reduces noise and produces robust results. LAD uses efficient algorithms to "discretize" the SELDI mass intensities by detecting "cutpoints" that discriminate in our case rapid from slow progressors. The discretization step may produce several binary features some of which may be redundant. LAD method employs combinatorial optimization algorithms to extract a "support set", a smallest (irredundant) subset of binary features, which can distinguish every pair of rapid and slow progressors in the SELDI dataset. "Patterns" are the key ingredients of LAD method. The "pattern generation" step uses the features in combination to produce a set of rules (combinatorial patterns) which can define homogenous subgroups of interest within the data. The simultaneous use of two or more features allows the identification of more complex rules which can be used for the precise classification of an observation. In "classification" step the method uses additional optimization techniques to generate a model consisting of a small number of patterns that can accurately classify patients as rapid or slow progressors. In the final step of LAD the model's performance is evaluated through cross-validation experiments.

Among other specific features of LAD which support its usefulness in biomedical informatics, LAD performs an exhaustive examination of the entire set of clinical, genomic, or proteomic features, without apriori excluding those which have either low statistical correlations with the outcome, or low expression levels. A novel and essential feature of LAD is its ability to discover not only potential biomarkers, but also potential "combinatorial biomarkers" (combinations of features) and its exhaustive search for underlying combinatorial patterns.

3 Results

The main goal of this study is to identify a small set of peaks to develop LAD models for the purpose of (i) classifying an individual observation as a rapid or slow progressor based on serum proteomic features and hence, (ii) predicting the progression of chronic kidney disease. The dataset for the 57 rapid progressors and 59 slow progressors consists of 5,751 intensities at intervals of 2 mass units derived from the SELDI-TOF data.

The LAD results described in the following sections were obtained with the use of "Ladoscope" (Lemaire 2005), a publicly available implementation of LAD, written and maintained by Pierre Lemaire. Each of the steps involved in the construction of a LAD classifier (discretization, feature selection, pattern generation, model selection, and classification), as described in this study, are available in Ladoscope software package.

3.1 Pre-processing

The SELDI-TOF data collected for this study contained many peaks, each of which potentially corresponds to the intensity level of a specific protein. In fact, many of these peaks are irrelevant for the recognition of a rapid progressor as opposed to a slow progressor. In order to obtain a classification model effectively and efficiently we applied a pre-processing procedure to retain only those relevant peaks distinguishing between rapid and slow progressors.

First we applied LAD method to generate high quality combinatorial patterns with characteristics (1) degree 1 – a single feature is used involved in the definition of the pattern; (2) homogeneity of at least 80% – the proportion of rapid (slow) progressor among all those patients covered by the pattern; and (3) prevalence of at least 80% – the proportion of rapid (slow) progressors covered by the pattern. We then retained the peaks participating in these patterns. After the application of this filtering procedure, we obtained a subset of 135 relevant peaks. Next we applied LAD method on the pre-processed data to generate higher degree LAD patterns to form an accurate classification model. In this set of 135 expression levels, we derived a minimal support set utilizing the intensities at 7 masses as listed in Table 2, which allows us to accurately distinguish the slow progressors from the rapid progressors. The statistical correlations of the selected features with GFR slope are also shown in the table.

It can be seen that many of the intensities correlate very poorly with slope. If all 5,751 features were ordered according to their decreasing correlations with the outcome (in absolute value), the attribute at m/z 2756 would appear as the feature with the correlation rank 16 (16th best correlation coefficient), but m/z 2018 would have rank 4,115 (see the last column of the Table). We emphasize these facts in order to point out that selecting the features only based one high correlations with the outcome may not necessarily lead to discovery of those which may have a significant predicting power when combined with other features.

SELDI Mass (m/z)	**Correlation Coefficient**	**Correlation Rank**
2018	0.039	4115
2756	0.260	16
2780	0.252	28
5266	0.065	3290
9940	0.194	348
11274	0.133	1565
11752	0.192	378

Table 2: Support set of SELDI-TOF masses whose intensities are used to create the LAD classification model

3.2 Classification Model

While an individual feature can partially predict an observation of being a rapid or slow progressor, the simultaneous use of two or more features allows for the definition of more complex rules (combinatorial pattern) that can be used for the precise classification of an observation. We applied the LAD method to the SELDI-TOF data consisting of 57 rapid and 59 slow progressors with the intensity levels of 7 masses presented in Table 2 to generate a classification model that can accurately define homogeneous subgroups of AASK samples with distinctive characteristics. The final LAD model classification model consists of three positive patterns (predicting rapid progression) and four negative patterns (predicting slow progression as presented in Figure 4 and the pattern characteristics including prevalence, homogeneity, and hazard ratio of each pattern are shown in Figure 5.

Patterns	Pattern defining conditions						
	m/z 2018	m/z 2756	m/z 2780	m/z 5266	m/z 9940	m/z 11274	m/z 11752
P1					< 0.575	> 0.055	
P2			< 3.835				> 2.78
P3	> 0.49				< 0.515		
N1			> 1.705		> 0.465		
N2				> 0.235		< 0.115	
N3		> 1.295			> 0.515		
N4				> 0.425			< 2.78

Figure 4: LAD Model for predicting rapid vs. slow progression in SELDI-TOF AASK data

Patterns	Pattern characteristics			
	Prevalence		Homogeneity	Hazard Ratio
	Positive	Negative		
P1	33 (57.89%)	10 (16.95%)	78.57%	2.42
P2	32 (56.14%)	8 (13.56%)	80%	2.43
P3	32 (56.14%)	9 (15.25%)	78.05%	2.34
N1	11 (19.30%)	39 (66.10%)	78%	2.57
N2	6 (10.53%)	31 (52.54%)	85.71%	2.39
N3	8 (14.04%)	35 (59.32%)	81.4%	2.48
N4	7 (12.28%)	31 (52.54%)	83.33%	2.3

Figure 5: LAD pattern characteristics

We remark that the patterns involved in the final classification model are high quality patterns since

- Each pattern has degree two (i.e., its definition involves at most two peaks) – small degree patterns (easy to interpret);
- On average positive patterns cover about 57% of the rapid progressors and negative patterns cover 58% of the slow progressors – high prevalence (each pattern covers several observations from the respected class);
- On average 79% of the observations covered by positive patterns are rapid progressors and 82% of the observations covered by negative patterns are slow progressors – high homogeneity (each positive (negative) pattern covers mostly rapid (slow) progressors and only a few slow (rapid) progressors).

The heatmap in Figure 6 shows the coverage of 116 AASK samples by the final LAD classification model.

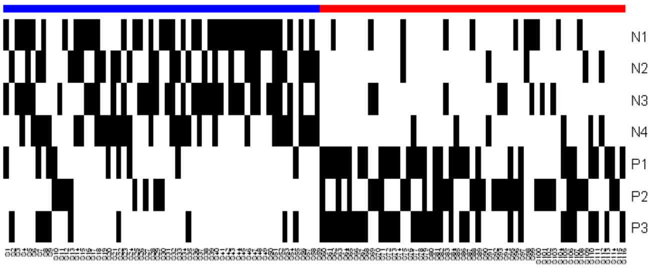

Figure 6: Heatmap of the LAD model
Blue: Slow progressors (negative class), Red: Fast progressors (positive class), **Black:** Pattern covers the observation

The overlapping black regions in the heatmap imply that an observation is covered by more than one pattern, and hence, only a few patterns are sufficient to separate the rapid and slow progressors. In the following section we evaluate the performance of prediction power of the patterns involved in the final LAD model.

3.3 Validation of the LAD Model

The performance of the final LAD classification model presented in 4 is evaluated through k-folding (10-folding in this case) cross validation technique: The SELDI-TOF data is randomly partitioned into $k = 10$ approximately equal parts; one of these subsets is designated as “test set”, a model is built on the remaining $k - 1 = 9$ subsets which form the “training dataset”, and then tested by classifying the cases in the test set using the model. This procedure is repeated $k = 10$ times, always taking another one of the ten parts in the role of the test set (re-randomizing the patients into 10 new subsets and repeat the procedure 9 additional times for a total of 100 tests). The average accuracy (proportion of correctly classified observations), sensitivity (proportion of correctly classified rapid progressors), and specificity (proportion of correctly classified slow progressors) are then reported as a quality measure of the proposed model in Table 3.

Accuracy	Sensitivity	Specificity	Hazard Ratio
$80.6 \pm 0.11\%$	$78.4 \pm 0.17\%$	$78.5 \pm 0.16\%$	2.72

Table 3: Cross validation of the LAD classification model

As can be seen, the LAD model predicts the rate of decline of kidney function among AASK samples with high sensitivity and specificity.

3.4 LAD Based Risk Scores

An observation whose measurements satisfy the defining conditions of a positive (negative) pattern, but do not satisfy the conditions of any of the negative (positive) patterns, can be easily "classified" as being rapid (slow) progressor. However, as can be seen in Table 5 and the heatmap in Figure 6, many patients satisfy the defining conditions both of some positive and of some negative patterns. The LAD model ultimately characterizes a patient as a rapid or slow progressor on the basis of a "discriminant" that takes into account all the patterns that cover the patient and employs a simple weighting procedure. The discriminant classifies patients as rapid (slow) progressors on the basis of their patterns in the model as follows: If the collection of all positive and negative patterns contains P positive and N negative patterns, and if a patient π satisfies the defining conditions of p of the positive and n of the negative patterns, then the discriminant value $\delta(\pi)$ of patient π is defined as $\delta(\pi) = (p/P) - (n/N)$, that is, the fraction of positive patterns that cover the patient minus the fraction of negative patterns that cover the patient. The patient π is then classified as being rapid or slow progressor, respectively, if the sign of $\delta\pi$ is positive or negative. Note that $|\delta(\pi)| <= 1$ and when $\delta(\pi) = 0$ the observation π is unclassified.

The discriminant can also be used as a risk score: the more positive the discriminant, the more likely the patient will be a rapid progressor. In order to a positive valued risk score to 116 AASK samples, we proceed as follows:

- Normalize the discriminant values by $(\delta(\pi) + 1)/2$ for all observations π in the dataset.
- Order the observations (AASK samples) according to their normalized discriminant values.
- Divide the set of observation into five equal size subsets (form the quintile of the 116 AASK samples).
- Calculate the average normalized discriminant values for each subset and associate it as the risk score of the AASK samples in the corresponding subset.

Table 4 depicts the risk scores assigned the AASK samples as well as the proportion of the rapid progressors in in each subset.

Risk Group	# of observations	% Rapid Progressors	Average Risk Score
1	23	0.00%	0.087
2	23	21.74%	0.275
3	23	47.83%	0.498
4	23	73.91%	0.697
5	24	100.00%	0.924

Table 4: LAD Risk Scores

As can be seen in Table 4, all of the patients in the highest and lowest risk groups were classified correctly by the LAD classification model and the risk scores are very close to the proportion of rapid progressors in each risk group. This significantly outperforms a similar classification of patients using proteinuria, the strongest traditional predictor of risk of progression in AASK, especially in identifying slow progressors. Table 5 and Figure 7 shows that the high levels of proteinuria accurately classify rapid progressors, misclassifying only a few patients, but the lowest proteinuria risk group misclassified 17% of the rapid progressors. Furthermore, the level of proteinuria in risk groups 1-3 varied from a UP/Cr ratio of 0.02 to 0.08 (probably on the order of 30-100 mg/day), levels that in the clinic would likely be indistinguishable, and 45% of patients in risk group 3 are rapid progressors.

Risk Group	# of observations	% Rapid Progressors	Average UP/UCr
1	23	16.7%	0.02
2	23	17.4%	0.03
3	23	47.8%	0.08
4	23	69.67%	0.31
5	24	95.7%	1.35

Table 5: UrineProtein/UrineCreatinine Risk Scores

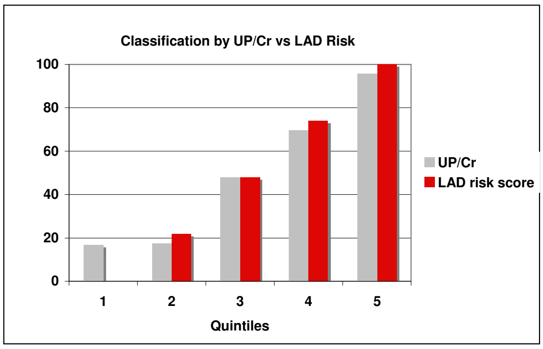

Figure 7: Proportion of patients classified as rapid progressors either by proteinuria or LAD risk score in risk quintile

We also generated receiver operating curves (ROC) as a measure of the effectiveness of the LAD discriminant at predicting GFR slope. As shown in Figure 8, the area under the curve is 0.899 (confidence interval 0.845 ± 0.953), which is highly significantly different from 0.5 (a value that would be found for a non-predictive test). A similar analysis using urine protein/creatinine ratio was also predictive but less robust (the area under the curve is 0.845 and confidence interval 0.774 ± 0.916), similar to the comparison in Figure 7.

4 Conclusions

In this study we present a pattern based classification approach to predict the rate of decline of kidney function in chronic kidney disease. We construct a classification model by applying the Logical Analysis of Data (LAD) methodology to the mass peak profiles of 57 rapid progressors and 59 slow progressors in African American Study of Chronic Kidney Disease with Hypertension (AASK) mass spectra data containing 5,751 serum proteomic features. The classification model has an accuracy of 80.6% obtained through

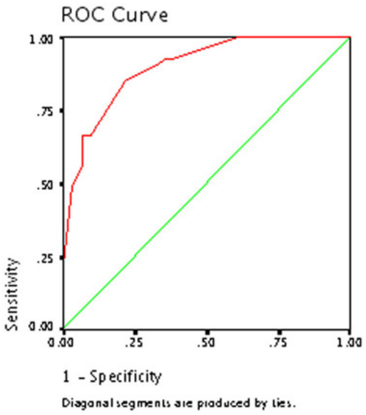

Figure 8: Proportion of patients classified as rapid progressors either by proteinuria or LAD risk score in risk quintile

ten times 10-folding cross validation experiments. In spite of significant predictive power of the LAD classification model, it contains 3 positive patterns (describing rapid progression) and 4 negative patterns (describing slow progression) and the patterns involved in the model are extremely simple (only involves two features in pattern description) and were developed using only 7 out of 5,751 peaks. The LAD discriminant is used as a risk score to identify the patients in different risk groups.

The ability to identify both progressors and non-progressors in groups of patients at risk for end-stage-renal-disease (ESRD) in itself is very important, in that it will allow us to focus efforts on those at greatest risk using the best current therapies. We may also be able to avoid complications of therapy if we can identify a truly low risk group that can just be observed (note that BP control to $< 130/80$ in AASK took almost 4 BP drugs). If an LAD risk score is associated with less than a 10% risk of progression, patients in that category could likely be observed rather than immediately treated with aggressive ACE/ARB therapy. While ACE and ARB are relatively benign, it should be noted that trials are underway with other drugs such as growth factor inhibitors (e.g., anti-TGF-b) that are likely to have much more aggressive side effect profiles, and it may be that even a higher risk would be justified before treatment. On the high risk end, a greater than 30-50% chance of progression would certainly seem to warrant treatment with ACE/ARB, but a cutoff will not be clear for newer therapies until trials are completed. Risk scores with a probability of progression in the 10-30/50% range may take individual judgment of risks/benefits before therapy. This may become increasingly important if new therapies such as anti-TGF-b drugs, which are likely to have a much more harmful side effect profile that ACE/ARB therapy, show some benefit for CKD progression. The ultimate goal of a future project motivated by this study is the identification of new protein targets that may suggest the new therapies that are desperately needed.

Acknowledgments

Special thanks to V. Anbalagan, W. Zhang, and the AASK Investigators. This research is supported by National Institutes of Health. Grant number: 5R21DK67468

References

Agodoa, L.; Appel, L.; Bakris, G.; Beck, G.; Bourgoignie, J.; Briggs, J.; Charleston, J.; Cheek, D.; Cleveland, W.; Douglas, J.; et al. 2001. Effect of ramipril vs amlodipine on renal outcomes in hypertensive nephrosclerosis: a randomized controlled trial. *Jama* 285:2719–2728.

Alexe, S.; Blackstone, E.; Hammer, P.; Ishwaran, H.; Lauer, M.; and Snader, C. 2003. Coronary risk prediction by logical analysis of data. *Annals of Operations Research* 119:15–42.

Alexe, G.; Alexe, S.; Liotta, L.; Petricoin, E.; Reiss, M.; and Hammer, P. 2004. Ovarian cancer detection by logical analysis of proteomic data. *Proteomics* 4:766–783.

Alexe, G.; Alexe, S.; Axelrod, D.; Hammer, P.; and Weissmann, D. 2005. Logical analysis of diffuse large B-cell lymphomas. *Artif Intell Med* 34:235–267.

Alexe, G.; Alexe, S.; Axelrod, D.; Bonates, T.; Lozina, I.; Reiss, M.; and Hammer, P. 2006a. Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Res* 8:R41.

Alexe, G.; Alexe, S.; Hammer, P.; and Vizvari, B. 2006b. Pattern-based feature selection in genomics and proteomics. *Annals of Operations Research* 148:189–201.

Appel, L.; Wright, J.; Greene, T.; Kusek, J.; Lewis, J.; X. Wang, X.; Lipkowitz, M.; others; and the AASK Collaborative Research Group. 2008. Long-term effects of renin-angiotensin system-blocking therapy and a low blood pressure goal on progression of hypertensive chronic kidney disease in african americans. *Arch Intern Med* 168(8):832–839.

Ball, G.; Mian, S.; Holding, F.; Allibone, R.; Lowe, J.; Ali, S.; Li, G.; McCardle, S.; Ellis, I.; Creaser, C.; et al. 2002. An integrated approach utilizing artificial neural networks and seldi mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics* 18:395–404.

Berg, U. 2006. Differences in decline in gfr with age between males and females. reference data on clearances of inulin and pah in potential kidney donors. *Nephrol Dial Transplant* 21:2577–2582.

Boros, E.; Hammer, P.; Ibaraki, T.; Kogan, A.; Mayoraz, E.; and Muchnik, I. 2000. An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering* 12:292–306.

Brenner, B.; Cooper, M.; de Zeeuw, D.; Keane, W.; Mitch, W.; Parving, H.; Remuzzi, G.; Snapinn, S.; Zhang, Z.; and Shahinfar, S. 2001. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *N Engl J Med* 345:861–869.

Crama, Y.; Ibaraki, T.; and Hammer, P. 1988. Cause-effect relationships and partially defined boolean functions. *Annals of Operations Research* 16:299–325.

DCCT. 1995. Effect of intensive therapy on the development and progression of diabetic nephropathy in the diabetes control and complications trial. The Diabetes Control and Complications (DCCT) Research Group. *Kidney Int* 47:1703–1720.

Fogarty, D.; Hanna, L.; Wantman, M.; Warram, J.; Krolewski, A.; and Rich, S. 2000. Segregation analysis of urinary albumin excretion in families with type 2 diabetes. *Diabetes* 49:1057–1063.

Hammer, P., and Bonates, T. 2006. An overview of logical analysis of data – from combinatorial optimization to medical applications. *Annals of Operations Research* 148:203–225.

Klahr, S.; Levey, A.; Beck, G.; Caggiula, A.; Hunsicker, L.; Kusek, J.; and Striker, G. 1994. The effects of dietary protein restriction and blood-pressure control on the progression of chronic renal disease. Modification of Diet in Renal Disease Study Group. *N Engl J Med* 330:877–884.

Krolewski, A.; Poznik, G.; Placha, G.; Canani, L.; Dunn, J.; Walker, W.; Smiles, A.; Krolewski, B.; Fogarty, D.; Moczulski, D.; et al. 2006. A genome-wide linkage scan for genes controlling variation in urinary albumin excretion in type II diabetes. *Kidney Int* 69:129–136.

Lauer, M.; Alexe, S.; Pothier-Snader, C.; Blackstone, E.; Ishwaran, H.; and Hammer, P. 2002. Use of the logical analysis of data method for assessing long-term mortality risk after exercise electrocardiography. *Circulation* 106:685–690.

Lemaire, P. 2005. The Ladoscope Gang: Tools for the Logical Analysis of Data. *http://www.kamick.org/lemaire/LAD/*.

Lewis, E.; Hunsicker, L.; Bain, R.; and Rohde, R. 1993. The effect of angiotensin-converting-enzyme inhibition on diabetic nephropathy. *N Engl J Med* 329:1456–1462.

Lewis, E.; Hunsicker, L.; Clarke, W.; Berl, T.; Pohl, M.; Lewis, J.; Ritz, E.; Atkins, R.; Rohde, R.; and Raz, I. 2001. Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. *N Engl J Med* 345:851–860.

Li, J.; Zhang, Z.; Rosenzweig, J.; Wang, Y.; and Chan, D. 2002. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem* 48:1296–1304.

Lindeman, R.; Tobin, J.; and Shock, N. 1985. Longitudinal studies on the rate of decline in renal function with age. *J Am Geriatr Soc* 33:278–285.

Lindeman, R. 1990. Overview: renal physiology and pathophysiology of aging. *Am J Kidney Dis* 16:275–282.

Murussi, M.; Gross, J.; and Silveiro, S. 2006. Glomerular filtration rate changes in normoalbuminuric and microalbuminuric type 2 diabetic patients and normal individuals – a 10-year follow-up. *J Diabetes Complications* 20:210–215.

Niki, P.; Panos, K.; and Christos, C. 2015. New targets for end-stage chronic kidney disease therapy. *The Journal of Critical Care Medicine* 1(3):9295.

Qu, Y.; Adam, B.; Yasui, Y.; Ward, M.; Cazares, L.; Schellhammer, P.; Feng, Z.; Semmes, O.; and Wright, G. 2002. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin Chem* 48:1835–1843.

Reddy, A.; Wang, H.; Yu, H.; Bonates, T.; Gulabani, V.; Azok, J.; Hoehn, G.; Hammer, P.; Baird, A.; and Li, K. 2008. Logical analysis of data (lad) model for the early diagnosis of acute ischemic stroke. *BMC Med Inform Decis Mak* 8:30.

Ruggenenti, P.; Perna, A.; Gherardi, G.; Garini, G.; Zoccali, C.; Salvadori, M.; Scolari, F.; Schena, F.; and Remuzzi, G. 1999. Renoprotective properties of ace-inhibition in non-diabetic nephropathies with non-nephrotic proteinuria. *Lancet* 354:359–364.

Wang, X.; Lewis, J.; Appel, L.; Cheek, D.; Contreras, G.; Faulkner, M.; Feldman, H.; Gassman, J.; Lea, J.; Kopple, J.; et al. 2006. Validation of creatinine-based estimates of gfr when evaluating risk factors in longitudinal studies of kidney disease. *J Am Soc Nephrol* 17:2900–2909.

Wright, J. J.; Bakris, G.; Greene, T.; Agodoa, L.; Appel, L.; Charleston, J.; Cheek, D.; Douglas-Baltimore, J.; Gassman, J.; Glassock, R.; et al. 2002. Effect of blood pressure lowering and antihypertensive drug class on progression of hypertensive kidney disease: results from the aask trial. *Jama* 288:2421–2431.