Pattern-based Classification and Survival Analysis of Chronic Kidney Disease

Munevver Mine Subasi[*][†] **Melissa Moreno**[‡] **Travaugh Bain**[§] **Megan M. Moreno**[¶]
Department of Mathematical Sciences, Florida Institute of Technology
150 W. University Blvd., Melbourne, FL 32901

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

Katherine C. Carroll[**]
Department of Biology
University of Florida
220 Bartham Hall
smallGainesville, FL 32611

Emily R. Cunningham[††]
Department of Mathematics
Potsdam SUNY
44 Pierrepont Avenue
Potsdam, NY 13676

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

Abstract

This study integrates the principles of pattern-based classification and Kaplan-Meier survival analysis to identify genes and clinical features associated with the rapid progression of chronic kidney disease. The methodology successfully determines the gene-gene survival interactions in the African-American Study of Chronic Kidney Disease with Hypertension (AASK) genomic dataset. The results obtained from this study serves as a basis for the future studies on comparison of the disease progression in white patients with that in African-American patients, both those with and those without apolipoprotein L1 (APOL1) high-risk variants.

1 Introduction

The main function of kidney is to remove excess water and waste products from blood. It also helps to regulate the levels of minerals such as sodium, calcium, and potassium in blood. Once suffers from chronic kidney disease (CKD) when kidney losses its function gradually, usually permanently. 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 severe renal failure and/or 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). However, progression rate of CKD is very heterogeneous (Lindeman, Tobin, and Shock 1985; Lindeman 1990; Hallan et al. 2006). While a few predictive factors for progression such as proteinuria have been detected, identification of those at risk to progress remains a significant problem. It has also been established that there are several therapies that can ameliorate the progression of renal disease including ACE inhibitors, blood pressure control, tight diabetes control and perhaps low protein diets; however, in trials examining these therapeutic modalities there remains a very significant risk of progression of renal disease in the subjects receiving optimal therapy (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. Since we use blood samples from AASK, for the proposed study, we shall discuss the problem in the context of the AASK results.

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 20 and 65 $mL/min/1.73m^2$. Other known causes of renal disease such as diabetes were exclusion criteria as was proteinuria $> 2.5gm/gm$ creatinine. Patients were randomized to the angiotensinogen converting enzyme inhibitor (ACEi) ramipril, the β-blocker (BB) metoprolol or the dihydropyridine calcium channel blocker (CCB) amlodipine, and to usual (mean arterial pressure (MAP 102-107) or low (MAP $<$ 92) blood pressure goals. The rationale for the treatment arms was that there was human and animal data suggesting that ACEi and CCB might slow progression of renal disease independent

[*]Email: msubasi@fit.edu
[†]Corresponding author.
[‡]Email: mmoreno2004@my.fit.edu
[§]Email: tbain2013@my.fit.edu
[¶]Email: mmoreno2010@my.fit.edu
[‖]Email: esubasi@fit.edu
[**]Email: katcarroll@ufl.edu
[††]Email: cunniner196@potsdam.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/

of their BP effects (Lewis et al. 1993; Hallan 1998), and there was data from observational and treatment studies that a lower BP might have beneficial effects (Klahr et al. 1994; Klag et al. 1997). Although other studies had attempted to achieve a $10mmHg$ MAP separation (Lewis et al. 2001; Hansson et al. 1998), AASK is the first major trial to actually achieve this goal. The primary outcome was rate of decline of GFR (GFR slope) based on iothalamate GFR studies at 6 month intervals, with a secondary clinical composite outcome of end stage renal disease (ESRD), a 25 ml/min or 50% drop in GFR from baseline (GFR event), or death.

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 CCB arm of the study was stopped early when interim analysis indicated that CCB was inferior to both BB and ACEi in patients with > 0.22 urine protein/creatinine ratio (about 300 mg proteinuria/24h) (Agodoa et al. 2001). The low BP goal of the study did not improve outcomes: there was no beneficial effect of low MAP on rate of progression of renal disease as defined by GFR slope or clinical composite outcomes (GFR events, end stage renal disease (ESRD) or death). Subsequently a similar result was found in the REIN trial (Ruggenenti et al. 1999). Studies in Type 2 diabetes have demonstrated a linear relation of achieved BP to renal outcomes (Bakris et al. 2003; Pohl et al. 2005); however, it should be noted that all the patients in these studies were treated to the same goal BP, so that rather than low BP being protective, the ability to achieve lower BPs may have defined a sub-population in these studies with low risks of disease progression. Despite lack of effect on renal outcomes in AASK, proteinuria was diminished by the lower BP goal. This finding is similar to that previously reported for diabetics (Lewis et al. 2001). Finally, a subgroup analysis in AASK did suggest that patients on a non-protective regimen (CCB) may have benefited from the low BP goal (Contreras et al. 2005). Most importantly in AASK, ACEi decreased the number of events as compared to both BB and CCB (Wright et al. 2002). These data for ACEi vs CCB are tabulated in Table 1 (risk reduction adjusted for baseline covariates) and were most dramatic for the hard outcomes, especially ESRD.

Ramipril vs. Amlodipine	% Risk Reduction	95% CI	p-value
GFR Event ESRD or Death	38%	14%-56%	0.004
GFR Event or ESRD	40%	14%-59%	0.006
ESRD or Death	49%	26%-65%	< 0.001
ESRD alone	59%	36%-74%	< 0.001

Table 1: Analysis of Clinical Composite Outcomes

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. Subasi et al. (2009) identified serum proteomic patterns: distinguishes fast progressors and slow progressors. Seldi-TOF mass spectra data containing 5731 serum proteomic features for 57 fast progressors and 59 slow progressors. Recently, Lipkowitz et al. (2013) examined effects of variants in gene encoding apolipoprotein L1 (APOL1) on progression of CKD and observed that renal risk variants in APOL1 were associated with the higher rates of ESRD and progression of chronic kidney disease in African-American patients as compared with white patients. Other recent studies include Rahman et al. (2013), where the effects of 2 antihypertensive drug dose schedules (PM dose and add-on dose) on nocturnal blood pressure vs. usual therapy (AM dose) in former participants were determined and Chen et al. (2015), where the longitudinal changes in hematocrit in hypertensive chronic kidney disease: results from the AASK was studied.

In this study we apply a pattern-based classification method and Kaplan Meier survival analysis method on AASK genomic and clinical data to identify clinical-genomic as well as gene-gene interactions to find putative prognostic markers for the progression of renal disease among AASK patients. Clinical and genomic features identified in our analysis will be used in a future study to Analysis of Data, to obtain comparison of the disease progression in white patients with that in African-American patients, both those with and those without apolipoprotein L1 (APOL1) high-risk variants.

2 Study Subjects and J48 Classification

Closer inspection of the data highlights the current dilemma: although there is a 30-60% decrease in the number of events with ACEiis still a residual event rate of $> 6\%/yr$ in the trial as a whole and $> 11\%/yr$ in subjects with urine protein/creatinine > 0.22, a mild degree of proteinuria of 200-300 mg/day (Figures 1 and 2). In addition it can be seen that the event rate is essentially constant throughout the 5 years of the trial, indicating that remaining patients are still at risk to progress. This finding is similar to that of other trials such as MDRD (Klahr et al. 1994; **?**), the Collaborative Study Group Trial (Lewis et al. 1993), RENAAL (Brenner et al. 2001) and IDNT (Lewis et al. 2001) .

There was significant heterogeneity of progression rate of renal disease in the AASK Trial as can be seen in Figure 3. 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 3

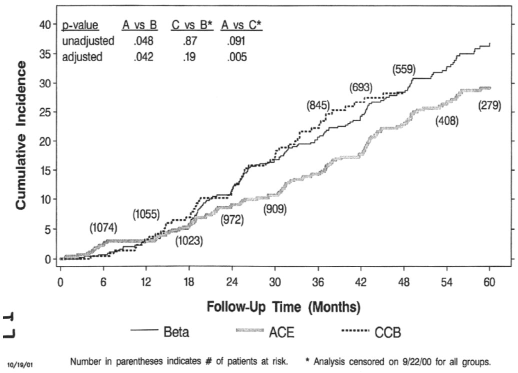


Figure 1: AASK Clinical Composite Events – all patients

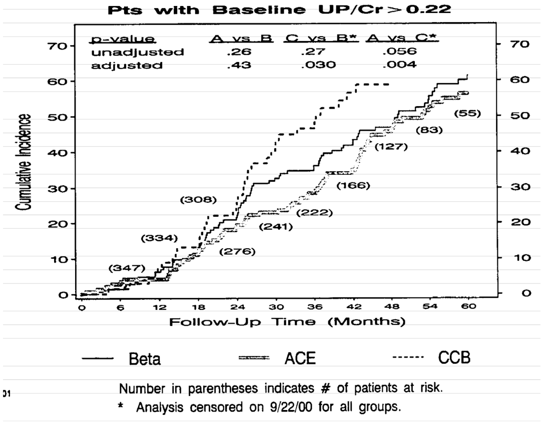


Figure 2: AASK Clinical Composite Events – proteinuria

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).

2.1 Pre-processing of AASK Data for Classification

An avenue that has not been carefully explored is a data mining approach to detect patterns of clinical features/serum protein expressions/SNPS that better identify the population at risk for progression of CKD. The goal of this section is to identify combinatorial patterns of clinical and SNPs that can accurately separate the progression of CKD. 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 glomerular filtration slope (GFR) of all AASK patients presented in Figure 3. The original AASK data contains 1,094 African-American patients with 88 clinical features and 130 single nucleotide polymorphism (SNPs). Before we have started our analysis we have removed all redundant features those with mostly missing values or no variation in the values as well as all patients with GFR values values are missing. This resulted in about 800 AASK patients with 77 clinical features and 113 SNPs for our analysis. In order to develop a classification model to identify the progression of CKD we have identified two “extreme” groups of patients where the disease progression is very small or very fast. Two sets of subjects were selected from the AASK study: “rapid progressors” (118-patients) and “slow progressors” (5-patients) based on the GFR histogram presented in Figure 4.

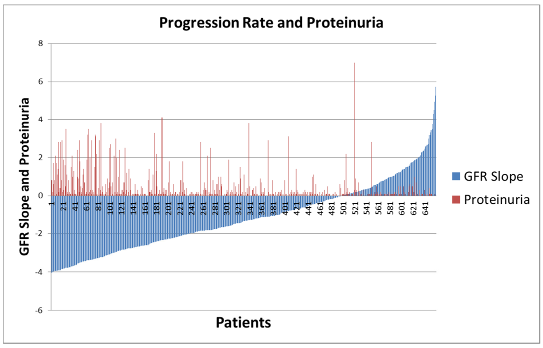


Figure 3: Patients stratified by GFR slope (blue bars) with degree of proteinuria superimposed (red spikes)

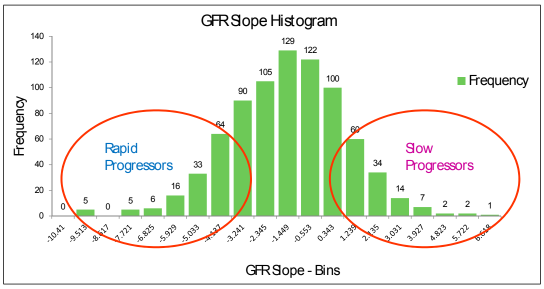


Figure 4: GFR Slope of AASK Patients

Table 2 describes the patient population for this pilot study.

	Rapid (n=138)	Slow (n=75)
Chronic Slope	-5.41 ± 1.36	2.11 ± 1.03
GFR	42.83 ± 13.25	52.30 ± 10.55
Proteinuria	1.12 ± 1.40	0.13 ± 0.20
Age	50.22 ± 11.94	52.52 ± 9.52
Weight (kg)	96.42 ± 22.42	87.52 ± 19.65
Height (cm)	171.69 ± 10.56	169.21 ± 10.80
BMI	32.69 ± 7.06	30.57 ± 6.09

Table 2: Baseline Characteristics of Study Population

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).

2.2 Feature Selection

The resulting AASK data consisting of 138 rapid progressors, 75 slow progressors, 77 clinical features, and 113 SNPs, is further investigated to remove any features 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 have applied a correlation-based feature selection procedure ((**?**)) to retain only those relevant features successfully distinguishing between rapid and slow progressors. Correlation-based feature selection method evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the outcome (rapid/slow progression) while having low intercorrelation are preferred. We have used WEKA ((Hall et al. 2009)), a commonly used open source data mining software, to perform the correlation-based feature selection procedure. Table 3 shows the features selected from the 10-fold stratified cross validation of the correlation-based feature subset selection procedure in WEKA.

Frequency in folds	% Frequency	Feature
9	90%	Perupherol base
10	100%	Proteinuria
10	100%	U.Protein/U.Creatinine
7	70%	GFR value at G1 visit
9	90%	CHGA-7
10	100%	CHGB-1
7	70%	FGB-G(-455)A

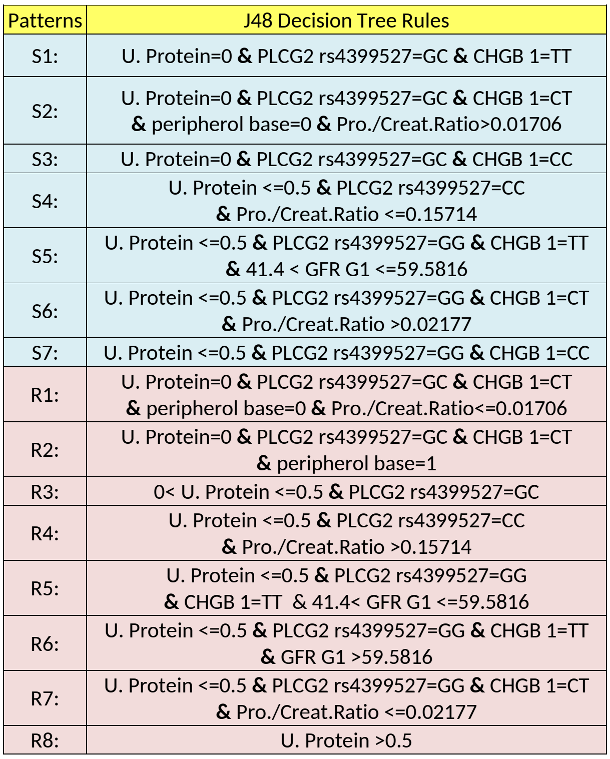
Table 3: Feature Selection - 10 fold stratified cross validation

We have observed that the SNP, CHGA-7, selected as a significant feature in feature selection step, contains GG for all AASK samples, except for one observation which is GA. Hence, we have removed CHGA-7 from further analysis. The resulting data contains 138 rapid progressors and 75 slow progressor three clinical features (proteinuria, urine-protein/urine-creatinine, GFR value at G1 vsit) and two SNPs (CHGB-1 and FGB-G(-455)A).

2.3 J48 Classification Model

In order to obtain a classification model consisting of combinatorial patterns of clinical features and SNPs we have used a powerful classification method, J48-decision tree method implemented in WEKA. J48 is an open source Java implementation of C4.5, an algorithm, that generates a decision tree developed by Quinlan (1993). Decision Tree is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

J48-decision tree procedure applied to the reduced AASK dataset from Section 2.2 provides us with the classification model shown in Figure 5. The classification model contains 7 patterns, S1-S7, for slow progressors and 8 patterns, R1-R8, for rapid progressors. Note that all patterns are combinatorial patterns of significant clinical features and SNPs obtained in Section 2.2.



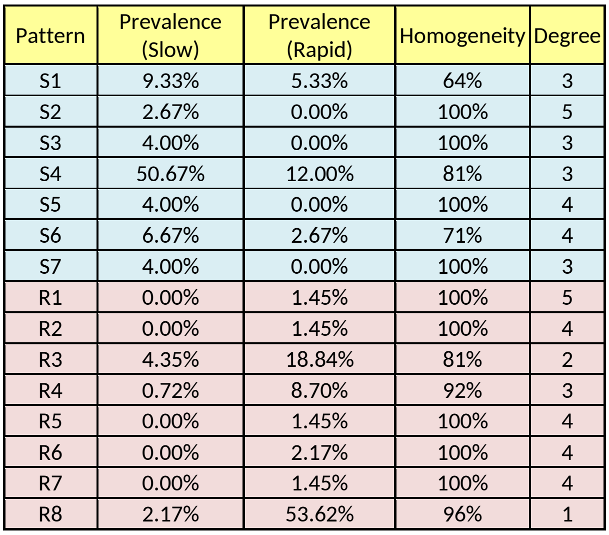
Patterns	J48 Decision Tree Rules
S1:	U. Protein=0 & PLCG2 rs4399527=GC & CHGB 1=TT
S2:	U. Protein=0 & PLCG2 rs4399527=GC & CHGB 1=CT & peripherol base=0 & Pro./Creat.Ratio>0.01706
S3:	U. Protein=0 & PLCG2 rs4399527=GC & CHGB 1=CC
S4:	U. Protein <=0.5 & PLCG2 rs4399527=CC & Pro./Creat.Ratio <=0.15714
S5:	U. Protein <=0.5 & PLCG2 rs4399527=GG & CHGB 1=TT & 41.4 < GFR G1 <=59.5816
S6:	U. Protein <=0.5 & PLCG2 rs4399527=GG & CHGB 1=CT & Pro./Creat.Ratio >0.02177
S7:	U. Protein <=0.5 & PLCG2 rs4399527=GG & CHGB 1=CC
R1:	U. Protein=0 & PLCG2 rs4399527=GC & CHGB 1=CT & peripherol base=0 & Pro./Creat.Ratio<=0.01706
R2:	U. Protein=0 & PLCG2 rs4399527=GC & CHGB 1=CT & peripherol base=1
R3:	0< U. Protein <=0.5 & PLCG2 rs4399527=GC
R4:	U. Protein <=0.5 & PLCG2 rs4399527=CC & Pro./Creat.Ratio >0.15714
R5:	U. Protein <=0.5 & PLCG2 rs4399527=GG & CHGB 1=TT & 41.4< GFR G1 <=59.5816
R6:	U. Protein <=0.5 & PLCG2 rs4399527=GG & CHGB 1=TT & GFR G1 >59.5816
R7:	U. Protein <=0.5 & PLCG2 rs4399527=GG & CHGB 1=CT & Pro./Creat.Ratio <=0.02177
R8:	U. Protein >0.5

Figure 5: Decision tree classification of AASK samples

The pattern characteristics such as prevalence (proportion of rapid(slow) samples covered by the patterns to total number of rapid(slow) samples), homogeneity (proportion of rapid(slow) samples covered by the pattern), and degree (number of conditions appear in the description of the pattern) of the J48 patterns are given in Figure 6.

2.4 Validation of the J48 Model

The performance of the final J48 classification model presented in Section 2.3 is evaluated through k-folding (10-folding in this case) cross validation technique: The AASK 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 $k = 30$ times, always taking another one of the ten



Pattern	Prevalence (Slow)	Prevalence (Rapid)	Homogeneity	Degree
S1	9.33%	5.33%	64%	3
S2	2.67%	0.00%	100%	5
S3	4.00%	0.00%	100%	3
S4	50.67%	12.00%	81%	3
S5	4.00%	0.00%	100%	4
S6	6.67%	2.67%	71%	4
S7	4.00%	0.00%	100%	3
R1	0.00%	1.45%	100%	5
R2	0.00%	1.45%	100%	4
R3	4.35%	18.84%	81%	2
R4	0.72%	8.70%	92%	3
R5	0.00%	1.45%	100%	4
R6	0.00%	2.17%	100%	4
R7	0.00%	1.45%	100%	4
R8	2.17%	53.62%	96%	1

Figure 6: Decision tree pattern characteristics

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 300 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 4.

Accuracy	**Sensitivity**	**Specificity**
77.5%	80.4%	74.6%

Table 4: Cross validation of the J48 classification model

As can be seen, the decision tree model predicts the rate of decline of kidney function among AASK samples with high sensitivity (true positive rate for rapid progressors) and specificity (true positive rate for slow progressors).

We have also generated receiver operating curves (ROC) as a measure of the effectiveness of the LAD discriminant at predicting GFR slope. As shown in Figure 7, the area under the curve is 0.918.

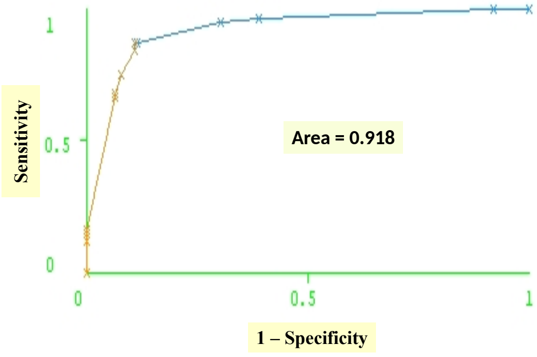


Figure 7: Receiver Operating Curves (ROC)

3 Survival Analysis

In this section we apply Kaplan-Meier survival analysis to AASK samples to determine the significant individual SNPs as well as the pairs of SNPs obtained in SNP-SNP analysis. Kaplan-Meier analysis is one of the best and commonly used survival methods to measure the fraction of subjects observing an event for a certain amount of time. In 1958, Edward L. Kaplan and Paul Meier collaborated to publish a seminal paper on how to deal with incomplete observations. Subsequently, the Kaplan-Meier curves and estimates of survival data have become a familiar way of dealing with differing survival times (times-to-event), especially when not all the subjects continue in the study. "Survival" times need not relate to actual survival with death being the event; the "event" may be any event of interest.

In AASK dataset we have had time-to-event data available for all patients, where an "event" is death, dialysis, and/or GFR event as shown in Tables 5 and 6.

	Death/Dialysis (DD)	Death/Dialysis/ GFR-Event (DDG)
Minimum	9.4	3.75
Mean± Std.Dev.	55.22 ± 14.61	49.59 ± 16.05
Median	52.4	49.5
Mode	65.2	48.6
Maximum	77.8	77.8

Table 5: Time-to-event information for AASK samples

	Death	Dialysis	G-Event	DD	DDG
# occurrences	13	101	137	114	187

Table 6: Time-to-event information for AASK samples

Initial pre-processing of the AASK data has included the removal of observations where time-to-event information is missing and/or most SNPs are missing we have obtained a dataset with 800 AASK patients and 113 SNPs. We have considered "Death/Dialysis/GFR-Event" as our time-to-event information and assumed that the data does not contain any censored samples.

3.1 Survival Analysis of AASK - Individual SNPs

When we have applied the Kaplan-Meier analysis on the data with 800 AASK patients and 113 SNPs, we have observed that the SNP MMP3-K45E(A/G) appears to be the most significant one. Of the 800 subjects analyzed, we were missing data for one patient for this SNP. The other patients appear to have similar curves. There were 97 patients homozygous with AA. They appear to have a slightly lower survival rate than the other patients. KCN-2 was the second most significant SNP in the analysis with a Kaplan-Meier p-value of 0.0001. Of the 800 subjects analyzed, 1 was homozygous with TT for KCN-2. 705 were homozygous with CC. 70 were CT and information was not available for 23 patients. Similarly, MMP2-C(-1306)T had the third most significant p-value. However, only one patient analyzed was homozygous with TT and two patients had missing data. 720

of the 800 subjects analyzed were homozygous for CC and appear to have survival rates similar to those with CT at this SNP. Based on our initial analysis we have further refined the dataset by removing all those SNPs which contain the same SNP value for all, except a few of the AASK samples.

Top 8 most significant SNPs (with p-value < 0.05) obtained as a result of the Kaplan-Meier survival analysis of 800 AASK samples is presented in Table 7.

SNP	Kaplan – Meier *p*-value
REN_4	0.006109041
SCNNIA ala663thr (G/A)	0.006985415
CYP11B2_rs1799998	0.003945369
GNAS FokI -/+ (T/C)	0.012371261
ACE A(-262)T	0.016874946
CYP3A4_rs2246709	0.0215345
PDE4D SNP42 A/G	0.039437293
AGTR1 G(-535)A	0.045274505

Table 7: Most significant SNPs predicting the survival of AASK patients

Kaplan-Meier survival curves of the 8 most significant SNPs are also presented in Figures 8-11.

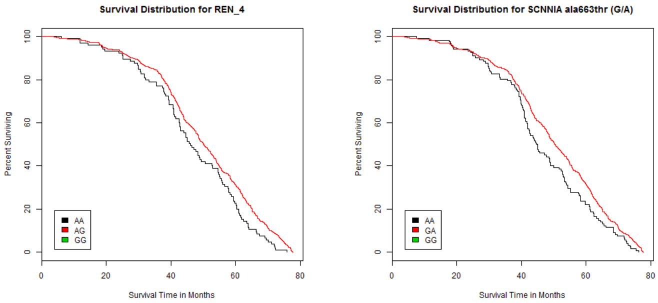


Figure 8: Top two most significant SNPs predicting the survival of AASK patients

3.2 Survival Analysis of AASK - SNP-SNP Analysis

Next we analyzed pairs of SNPs and looked for SNPs that appeared significant that previously did not appear individually significant. Multiple pairs of SNPs were found to be significant with p-values less than 0.00001. In Table 8 we present the top 10 most significant pairs of SNPs obtained by SNP-SNP survival analysis.

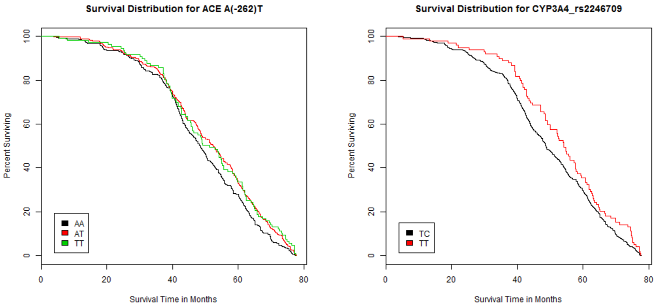


Figure 9: Third and fourth most significant SNPs predicting the survival of AASK patients

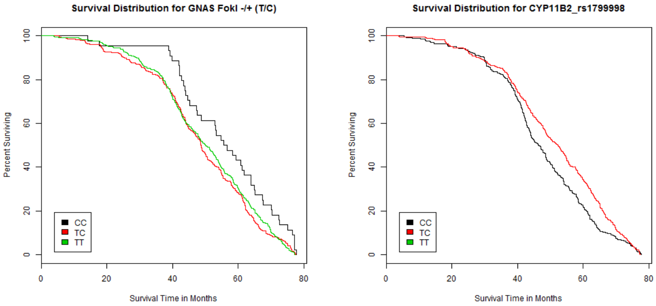


Figure 10: Fifth and sixth most significant SNPs predicting the survival of AASK patients

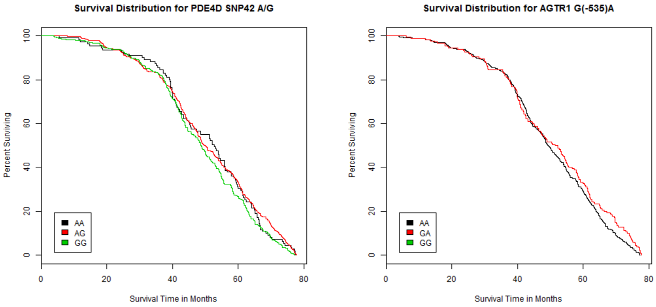
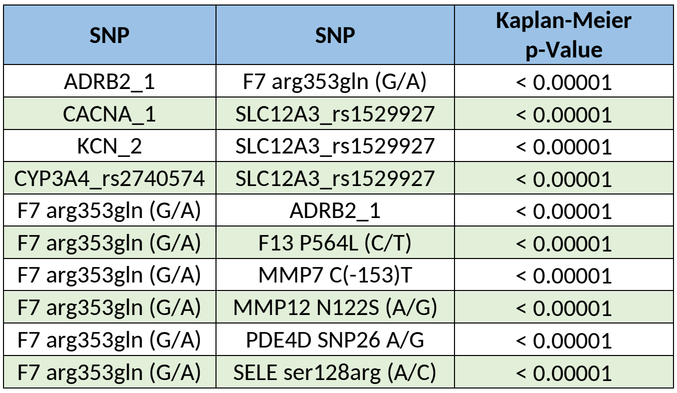


Figure 11: Seventh and eighth most significant SNPs predicting the survival of AASK patients

We remark that none of those SNPs obtained in SNP-SNP analysis were significant predicting the events when analyzed individually. F7 arg353gln (G/A) appears most often in Table 8, however, it had a p-value of 0.470 when Kaplan-Meier analysis was applied on individual SNPs. SLC12A3-rs1529927, the second most appearing SNP in pairwise Kaplan-Meier analysis, had a p-value of 0.729 when analyzed individually. Table 9 gives the the most significant SNPs listed in Table 8 together with their p-values obtained by Kaplan-Meier analysis on individual SNPs.



SNP	SNP	Kaplan-Meier p-Value
ADRB2_1	F7 arg353gln (G/A)	< 0.00001
CACNA_1	SLC12A3_rs1529927	< 0.00001
KCN_2	SLC12A3_rs1529927	< 0.00001
CYP3A4_rs2740574	SLC12A3_rs1529927	< 0.00001
F7 arg353gln (G/A)	ADRB2_1	< 0.00001
F7 arg353gln (G/A)	F13 P564L (C/T)	< 0.00001
F7 arg353gln (G/A)	MMP7 C(-153)T	< 0.00001
F7 arg353gln (G/A)	MMP12 N122S (A/G)	< 0.00001
F7 arg353gln (G/A)	PDE4D SNP26 A/G	< 0.00001
F7 arg353gln (G/A)	SELE ser128arg (A/C)	< 0.00001

Table 8: Top 10 pairs of SNPs predicting the survival of AASK patients

bf SNP	**p-value**
ADRB2-1	0.252606625
CACNA-1	0.426118598
F7 arg353gln (G/A)	0.46993914
CYP3A4-rs2740574	0.477507156
KCN-2	0.51776224
MMP7 C(-153)T	0.71102964
SLC12A3-rs1529927	0.728967435
MMP12 N122S (A/G)	0.747659507
SELE ser128arg (A/C)	0.801858508
PDE4D SNP26 A/G	0.92248453
F13 P564L (C/T)	0.986444372

Table 9: Most significant SNPs obtained in SNP-SNP analysis and their Kaplan-Meier p-values from individual SNP analysis

4 Conclusions

In this study we apply a pattern-based classification method and Kaplan Meier survival analysis method on AASK genomic and clinical data to identify clinical-genomic as well as gene-gene interactions to find putative prognostic markers for the progression of renal disease among AASK patients. We analyze the African-American Study of Chronic Kidney Disease (AASK) dataset and construct a decision-tree classification model consisting 7 combinatorial patterns of clinical features and SNPs for slow progressors and 8 combinatorial patterns of clinical features and SNPs for rapid progressors. The classification model uses only 4 clinical features and 3 SNPs and has an accuracy of 78.4% obtained through 30 times 10-folding cross validation experiments. We then apply Kaplan-Meier analysis to a dataset consisting of 800 AASK samples and 113 SNPs and identify significant individual SNPs as well as the pairs of SNPs that are obtained in SNP-SNP analysis.

The SNPs obtained in SNP-SNP analysis (p-value $<$ 0.00001) are not among the ones obtained as significant in individual SNP analysis (p-value < 0.05). We remark that none of those SNPs obtained in SNP-SNP analysis were significant predicting the events when analyzed individually. F7 arg353gln (G/A) appears most often in Table 8, however, it had a p-value of 0.470 when Kaplan-Meier analysis was applied on individual SNPs. SLC12A3-rs1529927, the second most appearing SNP in pairwise Kaplan-Meier analysis, had a p-value of 0.729 when analyzed individually. This shows the importance of considering combinatorial features (combinations of two or more SNPs).

We shall extend our analysis to obtain pattern-based survival analysis where we integrate the principles of classification algorithms with powerful Kaplan-Meier survival analysis. Clinical and genomic features identified in our classification as well as survival analysis will be used in a future study to obtain comparison of the disease progression in white patients with that in African-American patients, both those with and those without apolipoprotein L1 (APOL1) high-risk variants.

Acknowledgments

Katherine C. Carroll and Emily R. Cunningham involved in this project as REU participants of “Florida Tech REU Site: Research at the Intersection of Biology and Mathematics”, funded by National Science Foundation. Award Number: 1359341

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.

Bakris, G.; Weir, M.; Shanifar, S.; Zhang, Z.; Douglas, J.; van Dijk, D.; and Brenner, B. 2003. Effects of blood pressure level on progression of diabetic nephropathy: results from the RENAAL studyEffects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. *Arch Intern Med* 163:1555–1565.

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.

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.

Contreras, G.; Greene, T.; Agodoa, L.; Cheek, D.; Junco, G.; Dowie, D.; Lash, J.; Lipkowitz, M.; Miller, E.; Ojo, A.; et al. 2005. Blood pressure control, drug therapy, and kidney disease. *Hypertension* 46:44–50.

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.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutremann, P.; and Witten, I. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1).

Hallan, S.; Coresh, J.; Astor, B.; Asberg, A.; Powe, N.; Romundstad, S.; Hallan, H.; Lydersen, S.; and Holmen, J. 2006. International comparison of the relationship of chronic kidney disease prevalence and esrd risk. *J Am Soc Nephrols* 17:2275–2284.

Hallan, M. 1998. Calcium antagonists and renal disease. *Kidney Int* 54:1771–1784.

Hansson, L.; Zanchetti, A.; Carruthers, S.; Dahlof, B.; Elmfeldt, D.; Julius, S.; Menard, J.; Rahn, K.; Wedel, H.; and Westerling, S. 1998. Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the hypertension optimal treatment (hot) randomised trial. *Lancet* 351:1755–1762.

Klag, M.; Whelton, P.; Randall, B.; Neaton, J.; Brancati, F.; and Stamler, J. 1997. End-stage renal disease in african-american and white men 16-year mrfit findings. *Jama* 277:1293–1298.

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.

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.

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.

Parsa, A.; Kao, W.; Xie, D.; Astor, B.; Li, M.; Hsu, C.; Feldman, H.; Parekh, R.; Kusek, J.; Greene, T.; Fink, J.; Anderson, A.; Choi, M.; Wright, J.; Lash, J.; Freedman, B.; Ojo, A.; Winkler, C.; Raj, D.; Kopp, J.; He, J.; Jensvold, N.; Tao, K.; Lipkowitz, M.; Appel, L.; et al. 2013. APOL1 risk variants, race, and progression of chronic kidney disease. *N Engl J Med* 369(23):2183–2196.

Pohl, M.; Blumenthal, S.; Cordonnier, D.; De-Alvaro, F.; Deferrari, G.; Eisner, G.; Esmatjes, E.; Gilbert, R.; Hunsicker, L.; deFaria, J.; et al. 2005. Independent and additive impact of blood pressure control and angiotensin II receptor blockade on renal outcomes in the irbesartan diabetic nephropathy trial: clinical implications and limitations. *J Am Soc Nephrol* 16:3027–3037.

Quinlan, J. 1993. C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*.

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.

Subasi, M.; Lipkowitz, M.; Subasi, E.; Anbalagan, V.; Zhang, W.; Hammer, P.; Roboz, J.; et al. 2009. Logical analysis of data (LAD) applied to mass spectrometry data to predict rate of decline of kidney function. *DIMACS-RUTCOR Workshop on Boolean and Pseudo-Boolean Functions in Memory of Peter L. Hammer*.

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.