View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by D-Scholarship@Pitt

MODELS FOR EARLIER PROGNOSIS OF RENAL DECLINE IN POLYCYSTIC KIDNEY DISEASE (PKD) PATIENTS

by

Tiange Shi

BEng, Tianjin University of Science and Technology, China, 2016

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Tiange Shi

It was defended on

April 13, 2018

and approved by

Ada O Youk, PhD, Associate Professor of Biostatistics, Epidemiology, Clinical & Translational Science, Graduate School of Public Health, University of Pittsburgh

Andriy I. Bandos, PhD, Assistant Professor of Biostatistics and Radiology, Graduate School of Public Health, University of Pittsburgh

Thesis Director: Douglas Landsittel, PhD, Professor of Biomedical Informatics, Biostatistics, Clinical and Translational Science, School of Medicine, University of Pittsburgh

Copyright © by Tiange Shi

2018

Douglas Landsittel, PhD

MODELS FOR EARLIER PROGNOSIS OF RENAL DECLINE IN POLYCYSTIC KIDNEY DISEASE (PKD) PATIENTS

Tiange Shi, MS

University of Pittsburgh, 2018

ABSTRACT

Polycystic kidney disease (PKD) is a genetic condition that leads to increased formation and growth of kidney cysts, and thus may lead to rapid onset of end-stage renal disease (ESRD). In 2017, more than 15% adults in the US were estimated to have inherited PKD. About half of PKD patients require dialysis or renal replacement therapy by 60 years of age. Prior to these terminal outcomes, chronic kidney disease (CKD), which is defined as a progressive loss of kidney function, represents the primary outcome of interest for PKD patients. Treatments of CKD are currently being developed, bur need to be administered earlier in the process. Therefore, earlier prognosis of PKD patients at risk for renal decline provides an opportunity to prevent or delay the progression of ESRD and decrease morbidity and mortality.

In this study, CKD stage 3B with glomerular filtration rate (GFR) less than 45 ml/min was considered the endpoint of highest clinical interest since stage 3B is early enough to identify patients before rapid decline, but late enough to represent a clinically meaningful outcome. We evaluated earlier prognostic ability of factors available at birth for CKD stage 3B among currently healthy PKD patients; use of only factors available at birth is a novel approach and could lead to early identification of PKD patients who subsequently experience clinical outcomes (e.g. later stages of CKD or ESRD).

Training data were collected from the Consortium for Radiologic Imaging Studies of Chronic Kidney Disease (CRISP). Multivariable logistic regression was initially employed to predict renal decline. A pruned classification tree model showed similar prognostic ability as logistic regression based on overlapping 10-fold cross validation AUC confidence intervals. Random forests, however, showed significant improvement in prognostic ability.

This study also validated results using a completed clinical trial of similar PKD patients (the HALT Progression of Polycystic Kidney Disease Study). Both CRISP cross validation and HALT validation results agreed on the best model (random forests) for prognostic ability.

In terms of public health significance, random forests could help estimate the probability of PKD patients reaching renal failure at given age, and thus inform prevention efforts.

TABLE OF CONTENTS

PREFACE ..... **IX**

1.1 POLYCYSTIC KIDNEY DISEASE **1**

1.2 CHRONIC KIDNEY DISEASE **2**

1.3 COMMON PROGNOSTIC MODELS FOR BINARY OUTCOMES **4**

1.4 RANDOM FORESTS **6**

2.1 DESCRIPTION OF DATASETS **7**

2.1.1 CRISP dataset **7**

2.1.2 HALT dataset **9**

2.2 METHODOLOGY **10**

2.2.1 Logistic regression **10**

2.2.2 Classification trees **11**

2.2.3 Random forest **14**

2.2.4 Variable importance measures **15**

2.2.5 Differences between random forests and bootstrap **16**

2.2.6 Development of random forests **17**

2.2.7 Validation and assessment of prognostic accuracy **18**

3.1 DATA DESCRIPTION **19**

3.2 RESULTS FROM LOGISTIC REGRESSION **22**

3.3 RESULTS FROM CLASSIFICATION TREES **26**

3.4 RESULTS FROM RANDOM FORESTS **30**

3.5 MODEL VALIDATION **31**

3.6 ADJUSTED COMPARISON BETWEEN THREE MODELS 33

APPENDIX A : STATA CODES 37

APPENDIX B : R CODES 44

BIBLIOGRAPHY 67

LIST OF TABLES

Table 1. Demographic data descriptive summaries for CRISP 19

Table 2. Gene mutation data descriptive summaries for CRISP 20

Table 3. Demographic data descriptive summaries for HALT 21

Table 4. Gene mutation data descriptive summaries for HALT 22

Table 5. Results of the unadjusted logistic model for demographics 23

Table 6. Results of the logistic model for each gene coding adjusting for demographics 23

Table 7. 10-fold cross validation AUCs for logistic models adjusting for different gene mutation 24

Table 8. Coefficients of the best logistic model 25

Table 9. Summary of AUC results of logistic regression and tree model 30

Table 10. Importance of variables 30

Table 11. Summary of AUC results of validation 31

Table 12. Summary table for model training and validation AUC 33

Table 13. AUC comparison between three models using PKD1 vs. PKD2/NMD gene mutation 34

LIST OF FIGURES

Figure 1. Splitting node in classification tree 12

Figure 2. 10-fold cross validation ROC for logistic models adjusting for different gene mutation codings 24

Figure 3. Individual ROC curves for each main effect in the best logistic model..... 26

Figure 4. Tree model without pruning classifying CKD stage 3B 27

Figure 5. Cross-validation relative error for each complex parameter 28

Figure 6. Pruned tree for predicting stage CKD 3B..... 28

Figure 7. 10-fold cross validation ROC for pruned tree model 29

Figure 8. Pruned tree model with age alone 29

Figure 9. ROC for random forest model..... 31

Figure 10. Validation ROC curves..... 32

PREFACE

I would like to express my genuine gratitude to Professor Douglas Landsittel, my academic advisor, who introduced the topic of my thesis and has always been encouraging and inspiring throughout my graduate study and research. I could not have finished my current work without his strong support. I am also very grateful to Professor Ada Youk and Professor Andriy Bandos not only for sharing their valuable insights and illuminating comments on this thesis.

I would also like to thank my family, especially my parents Shi, Lin and Wang, Ping for supporting my back with love and understanding during my time pursuing my master's degree at University of Pittsburgh. If any glory, it should be dedicated to them as well.

1.0 INTRODUCTION

1.1 POLYCYSTIC KIDNEY DISEASE

Autosomal dominant polycystic kidney disease (ADPKD) is one of the most widespread genetic cystic kidney disorders; ADPKD often leads to premature development of end-stage renal disease (ESRD). [1-3] More specifically, nonreplicable cysts expansion and development inside bilateral kidneys leads to progressive renal enlargement, fibrosis and parenchyma destruction. These physiological changes lead to seriously impaired quality of life for around 600,000 Americans ADPKD patients annually. [4] Among those complications, ADPKD leads to hypertension, cerebral aneurysms, cardiovascular disease and hepatic cysts development. [5, 6] More than 60% of ADPKD patients are reported to suffer from hypertension or gross hematuria [7-9], and about 80% of adults with ADPKD are also detected with polycystic liver disease. [10]

Another important characteristic of ADPKD is genetic polymorphisms. Studies have shown that at least two genes are associated with this disease. [11] About 80-85% of ADPKD cases were caused by mutations to polycystic kidney disease type 1 gene (PKD1), and mutations to PKD2 account for most of the remaining cases. [11,12] Heyer et al [13] showed that patients with PKD1 mutations tend to have more severe disease than PKD2; truncating (i.e., nonsense mutations, frameshift mutations, splicing mutations, and large rearrangements) PKD1 mutations were associated with worse renal function loss compared to non-truncating PKD1 mutations. From

mutation strength level, the truncating PKD1 population was set as mutation strength group 1 (MSG1); the non-truncating PKD1 population was then further divided into strong mutation strength group (MSG2) and weak mutation group (MSG3). Heyer et al [13] showed that patients within MSG3 had significantly higher (i.e., better) glomerular filtration rate (GFR) than truncating PKD1 mutation group (MSG1).

The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) was funded by the National Institute of Diabetes and Digestive and Kidney Disorders to identify biomarkers of disease progression in patients with ADPKD. Previous publications from CRISP have shown that higher baseline total kidney volume (TKV) and lower renal blood flow (RBF) are positively associated with disease progression in ADPKD. [14, 15] Other studies also showed that serum or urine biomarkers, including monocyte chemoattractant protein-1 (MCP) and blood urea nitrogen (BUN), having a role predicting the renal failure in PKD patients. [16, 17] This study will contribute to the literature by evaluating prognostic ability of factors available at birth, which can then possibly allow for earlier intervention.

1.2 CHRONIC KIDNEY DISEASE

Chronic kidney disease (CKD), which is defined as a progressive loss of kidney function, is the primary outcome of interest for PKD patients. CKD is also associated with hypertension, diabetes, cardiovascular disease, nephritis, and other complications. [18] CKD poses a huge threat to global public health. CKD has a global prevalence of about 10% [19] and more than 30 million people in the US suffer from CKD. [20]

A number of different strategies can be used to diagnose and classify CKD. One approach identifies CKD using any one of three characteristics: cause, glomerular filtration rate (GFR) and albuminuria. [21] Identifying cause is emphasized because of its fundamental importance in predicting outcome and guiding choice of cause-specific treatments. Based on the GFR, CKD can be classified into one of 5 stages: 1) fully functional kidneys with GFR over 90 ml/min, 2) somewhat reduced but still normal levels of preserved renal function with GFR between 60-90 ml/min, 3) mild to moderate reduction in renal function with GFR between 30-60 ml/min, 4) severe reduction in renal function with GFR between 15 and 30, or 5) renal failure with GFR below 15. Stage 3 CKD may also be divided into stage 3A (mild to moderate reduction) with GFR between 45 and 60 or stage 3B (moderate, approaching severe reduction) with GFR between 30 and 45. [22, 23] The actual GFR can be assessed through measuring iothalamate clearance, but is usually impractical to assess on regular basis; therefore, most studies and clinical applications use serum creatinine (with age, sex and race) to calculate estimated GFR (eGFR). [24]

While there are a few strategies for treating or slowing the effect of CKD, such as the use of blood pressure medications. [25, 26] Prevention and treatment to CKD are severely challenged by the inability to identify those patients who decline more quickly, versus those whose GFR remains relatively constant (or declines slowly) over time. Notable declines in GFR measurements and detection of albuminuria typically do not happen until the patient is already on a course for rapid decline. This narrows the valuable time window for effective clinical intervention. Greene et al [27] suggested that use of the 40% eGFR declines was an appropriate strategy to reduce sample size. However, the complexity of eGFR trajectories, which can be highly variable and remain constant until soon before rapidly falling toward renal failure, introduces significant

challenges for interventions. Thus, developing prognostic models for renal decline that are sufficiently accurate across the population remains an imperative goal for public health.

Grantham [28] showed that kidney enlargement resulting from the expansion of cysts in patients with PKD is associated with the decline of renal function. In terms of risk factors associated with progression of cyst growth and renal decline, Chapman [29] showed that baseline total kidney volume (TKV) predicts CKD development almost a decade later, thus motivating use of TKV as a prognostic biomarker in PKD. Based on these results, new diagnostic approaches using machine learning algorithms were introduced to predict CKD. In a previous study, artificial neural networks [30] were shown to produce accurate prognosis of renal failure in PKD patients. In another analysis using tree models [31], optimally pruned tree models used only two variables, baseline eGFR and baseline height adjusted TKV, to accurately predict CKD status. This current study, in addition to focusing on factors available at birth, is the first to also validate results using a completed clinical trial of similar PKD patients.

1.3 COMMON PROGNOSTIC MODELS FOR BINARY OUTCOMES

Both logistic regression and classification tree model are commonly used in disease prognosis studies with binary outcomes. Logistic regression was proposed as a method for binary outcomes in the late 1960s. [32] It models the probability of an outcome of interest by several predictive factors in an equation of the form

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$$

where p is the probability of the outcome, β_0 is the intercept, $\beta_1, \ldots, \beta_i$ are the coefficients of predictors $x_1, \ldots, x_i$.

Logistic model is easy to interpret; One can easily convert β coefficients to the corresponding odds ratios and interpret the magnitude of importance of the predictors. Thus, logistic regression is probably the most popular statistical technique used to describe relationships between independent variables and a dichotomous dependent variable. However, logistic regression forces all the predictive factors in a linear function to the log odds of the outcome and does not implicitly account for non-linearity or interactions. These non-linearity and interactions need to be explicitly specified, and the linearity assumption of logistic regression is hard to be met as well.

Classification trees is another method for classification problems, where data are recursively split into two groups until the final subsets (or terminal nodes) are too small to split further (usually set at n<10) or data are perfectly classified. [33] Classification tree algorithm has been broadly applied for classifying binary outcomes in numerous areas of medicine. [34] This approach is more flexible and can implicitly model non-linearity and interactions in the data. Results may therefore be superior for nonlinear classifications and complex sample structures. Tree models can be pruned by deleting non-significant splits (from the bottom of the tree working up) and thus avoid over-fitting the data. [35]

However, the classification trees algorithm is known to be unstable. Tree models can produce drastically different results from training datasets that differ just slightly. [36] This instability undermines the objective of extracting knowledge from the trees.

1.4 RANDOM FORESTS

Breiman proposed bagging of classification trees in 1996 [37], in which successive trees are constructed independently using a bootstrap sample of the dataset and a simple majority vote is taken for prediction. Later in 2001, Breiman proposed random forests [38], which add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables, while, in a random forest, each node is split using the best split among a subset of predictors randomly chosen at that node. This strategy turns out to preform very well compared to many other classifiers, including discriminant analysis and other machine learning methods, and is robust against overfitting [38].

In this thesis, we used multivariable logistic regression, classification trees, and random forests to produce prognostic models for identifying kidney decline earlier in the patients' lifetime (as defined by CKD Stage 3B), and to compare their accuracy. Receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) will also be employed to assess the classification ability of methods proposed in this study. Results will provide both an illustration of the strengths and limitations of these methods, and potentially new findings for the CRISP data, which may have important implications for prognosis and earlier treatment of CKD. Methods also focus on factors available only at birth to assess whether risk stratification is possible very early in life.

2.0 MATERIAL AND ANALYSIS

2.1 DESCRIPTION OF DATASETS

In this thesis, we used datasets one observational study and one completed clinical trial (where treatment had no significant effect) of similar PKD patients. Training data used in this study were from (the observational study) Consortium for Radiologic Imaging Studies of Chronic Kidney Disease (CRISP). Validation data were obtained from HALT Progression of Polycystic Kidney Disease Study (HALT).

2.1.1 CRISP dataset

The PKD dataset used in this study is taken from the Consortium for Radiologic Imaging Studies of PKD (CRISP). [10] This prospective cohort study recruited 241 adults with ADPKD and preserved kidney function and followed them up during 2001 to 2016. GFR or estimated GFR (eGFR) were used to detect whether a participant reached a renal insufficient endpoint. Stage 3B CKD where GFR less than 45 ml/min is considered as the endpoint of interest. GFR was detected using corrected iothalamate clearance for CRISP and estimated GFR using the CKD-EPI equations in HALT (since corrected iothalamate clearance was not available in HALT); previous data (not shown here) have shown a high level of agreement. [15] For the CRISP training data, ten participants were found to have stage 3B CKD at baseline and were excluded from the study. Patients found without CKD and with follow-up less than 10 years were also excluded, since it could not be determined if they had the outcome at a given time point.

Demographic information, including gender, age, race, birthweight and BMI, were collected at baseline. Clinical information, MR determined TKV, liver cyst volume (LCV), urinary monocyte chemo attractant protein (MCP), renal blood flow (RBF) and blood Urea Nitrogen (BUN) were collected at clinic visits (initially annually for the first years after baseline) during the 15-year follow-up (with a maximum follow-up of 14.2 years). Polycystic kidney disease genotype and corresponding mutation strength were also collected during the study.

This thesis focuses entirely on CKD prognosis using factors available at birth. Predictor variables were gender, race, gene mutation and different measures of mutation strength. CRISP recorded different gene mutation strength information in several different ways: (1) gene mutation (PKD1 mutation, PKD2 mutation and no mutation detected (NMD)); (2) mutation types (truncated mutation versus non-truncated mutation); (3) mutation strength groups (MSG), where MSG1 corresponds to truncating PKD1 mutations; MSG2 and MSG3 are strong and weak mutations, respectively, divided from non-truncating PKD1 and PKD2 mutation populations. [13] (4) Semi-continuous mutation strength score (SCMSS), which divided mutation strengths into a 6-point scale, where 0 is for no mutation, 1 as the mildest and 5 as the worst. In addition to these factors which are available at birth, age will also be included in the model to account for where participant age at baseline measurements.

Descriptive statistics for all these predictor variables were shown in Table 1 and 2 in Chapter 3.1. For outcome of interest, number (percent) of observations reaching CKD stage 3B were shown in the first row of each table. Descriptive statistics were described using the mean (standard deviation) for continuous variables, and count (percent) for categorical variables. P-values for age and birthweight were calculated using Wilcoxon rank sum test; chi-square test was used for categorical variables.

2.1.2 HALT dataset

The HALT Progression of Polycystic Kidney Disease (HALT) study is the first prospective, randomized clinical interventional study for adults with ADPKD. [39] Two simultaneous multicenter clinical trials (study A and study B) were conducted to test the efficacy of interruption of the renin-angiotestin-aldosterone system (RAAS) on the progression of cystic disease and the decline in renal function in ADPKD. [40] Study A investigated treatment effects on patients with early ADPKD defined by GFR greater than 60 ml/min, where the change of TKV served as the primary outcome and eGFR as the secondary outcome measure. The objective of study B was to investigate the treatment efficacy on the time to a 50% reduction of baseline eGFR, ESRD or death, among hypertensive individuals with moderately advanced PKD defined by GFR 25-60 ml/min. [39] Because the study population of HALT A was more similar to CRISP, only data from HALT study A were used as validation. For purpose of this thesis, the actual intervention in HALT A (which was not significant) was ignored.

Study A enrolled individuals aged 15 to 49 diagnosed with ADPKD and a eGFR greater than 60 ml/min. The eGFR was collected twice a year, and stage 3B CKD where eGFR less than 45 ml/min was considered the endpoint of interest. Over the whole cohort, median follow-up time was 6 years. Patients found without CKD with follow-up less than 5 years were excluded, because the length of follow-up was not sufficient to determine the outcome. Descriptive statistics for demographic and gene mutation variables were shown in Table 3 and 4. P-values for age were calculated using Wilcoxon rank sum test; chi-square test was used for categorical variables.

2.2 METHODOLOGY

2.2.1 Logistic regression

Logistic regression is the most standard method for predicting a binary outcome. It extends the simple linear regression by applying logistic function. In our study, we will use multivariable logistic regression model to predict the CKD stage 3B in patients. The model is defined as:

$$\ln\left(\frac{p}{1-p}\right) = X\beta$$

where p is the probability for a patient reaching the given stage of renal decline, with covariate matrix X and parameter vector β. Besides the genotype and mutation strength [43], relative studies also indicated that lower birthweight [41] and younger age at baseline [42] are correlated with having CKD outcome. Although age is obviously not a variable "at birth", it was included to account for how far the participant was in time when outcome data were collected. This thesis thus used birthweight, race, gender, age and gene mutation scores as predictor variables. Collinearity was assessed with variance inflation factors (VIF) and we deleted variables that induced collinearity to obtain the final model. Logistic models were compared using likelihood ratio test (LRT), Bayes Information Criteria (BIC) and 10-fold cross validation AUC to determine the best model.

To test whether factors available at birth add significantly to current age in predicting renal decline, we first fit a logistic model with age only versus age and other factors available at birth (gender, race and birthweight) for observations with birthweight records.

Because mutation strength information collected by both studies was highly correlated and cannot be fitted into the same logistic regression model, another goal of this study was to determine

the best coding of gene mutation. According to the characteristics of each gene mutation information, we proposed four ways of coding: (1) PKD1 versus PKD2 or NMD; (2) PKD1-truncating versus PKD1-nontruncating versus PKD2 or NMD; (3) PKD1 within MSG1 versus PKD1 within MSG2 and 3 versus PKD2 or NMD. (4) SCMSS and gene mutation (PKD1 versus PKD2), with both main effects included in the given model. Because almost half (42.42%) observations were missing birthweight in CRISP data, analyses were first performed without birthweight. To select the best way of coding gene mutation strength, we fitted four multivariable logistic models adjusted for different gene mutation codings. The LRT for each gene mutation coding was performed to test the significance of prognostic ability and models were compared using Bayes Information Criteria (BIC). Model with the smallest BIC was selected to be the best the model, and corresponding gene mutation would be the best coding. Results were shown in Table 6 in Chapter 3.

2.2.2 Classification trees

Classification trees predict the outcome for a given subject in the validation set as the most commonly occurring class of training observations in the region to which it belongs based on recursive partitioning of the data. The proportion of subjects with the outcome in a particular terminal node region (based on the training observations) defines the predicted proportion (or probability) for that subject. In terms of the model fitting process, data are recursively split into two groups (shown in Figure 1) until the final subsets meet with stopping rules, which are achieving final subsets are too small to split further ($n<10$ in our cases) or data that are perfectly classified.

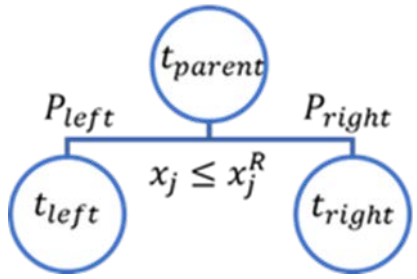

Figure 1. Splitting node in classification tree

Consider data from distribution (Y, X) where $Y \in \{1,\ 2,\ \ldots, M\}$ is the class label and $X = (x_1,\ \ldots, x_G) \in \mathbb{R}^G$ is the G-dimensional covariates. To fully grow the classification tree, at each split node, the tree model uses a factor x_j to split parent data, t_{parent}, into two child datasets, t_{left} and t_{right} with the probabilities of p_{left} and p_{right}. After each split, the homogeneity of the child dataset will increase, and the impurity of child dataset will decrease compared to the parent dataset. To measure the impurity, we used GINI function $i(t)$:

$$i(t) = \sum_{m=1}^{M} P_m(1 - P_m)$$

To identify the best splitting rule of each node, the corresponding x_j value can be calculated by maximizing the change of GINI at each split:

$$\underset{x_j \leq x_j^R,\, j=1,\ldots,M}{\arg\max}\ [\Delta i(t)] = \underset{x_j \leq x_j^R,\, j=1,\ldots,M}{\arg\max}\ [i(t_p) - P_l i(t_l) - P_r i(t_r)]$$

The classification tree will continue growing until the stop rules are reached.

To avoid overfitting issue, full-grown tree needs to be pruned. In this thesis, we used minimal cost complexity pruning method. More specifically, let $A_1, A_2, \ldots, A_n$ denote all nodes in full-grown tree while $R(A_i)$ denote the risk of corresponding node A_i. $|T|$, total number of split nodes, describes the complexity of a classification tree T. The risk of a classification tree T is then calculated by summation of all the risk of split nodes in T:

$$R(T) = \sum_{i}^{n} P(A_i)\, R(A_i)$$

When pruning the tree, the new risk of the pruned tree is introduced:

$$R_\alpha(T) = R(T) + \alpha|T|$$

where α denotes the complexity parameter, which seeks to balance the prediction error rate and number of nodes in the tree model. The new risk for the pruned tree is penalized for the misclassification risk $R(T)$ and the total number of terminal nodes $|T|$ based on the complex parameter α. The tree is pruned with complexity parameter that minimized $R_\alpha(T)$, i.e.,

$$R_\alpha(T(\alpha)) = \min_{T \leq T_{max}} R_\alpha(T)$$

The pruning processes can be interpreted as follows: starting from the bottom of a classification tree, the terminal nodes could remain in the model when the decrease of misclassification risk is greater than α times of the change of tree complexity $|T|$.

Classification tree models have a few advantages comparing to other classical algorithms e.g. linear regression, logistic regression: Tree models are easy to interpret, even more so than generalized linear models. Presumably, classification trees more closely reflect human decision-making than do the regression and other classification approaches. Trees can be displayed graphically and are easily interpreted by non-experts. Trees can easily handle qualitative predictors without the need to create dummy variables.

There are also disadvantages of tree models: Tree models may not have the same level of predictive accuracy as some more complex machine learning approaches. Trees can be very unstable. In other words, a small change in the data can cause a large change in the final estimated tree.

2.2.3 Random forest

A random forest model, as proposed by Leo Breiman in 2001, is a fast, often highly accurate, noise resistant classification method. [44] Random forests use an ensemble of classification trees, where each tree is built using 1) a bootstrap sample of the data, and 2) a random selection of predictors at each node. Hence, the random forest includes a combination of bootstrap sampling and random selection of variables that potentially differs at each node of the trees.

The bootstrap aggregating (or bagging) is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method. Use of these resampling approaches was first introduced by Breiman [37,45] to stabilize the relatively unstable tree algorithm procedure. The bagging algorithms in a classification setting can be described as follows:

Consider a learning set $\mathcal{L}$ that consists of data $\{(y_n, x_n), n = 1, \dots, N\}$ where the y's are class label of outcome. Assume we have a procedure for using this learning set to form a predictor $\varphi(x, \mathcal{L})$; if the input is x, we predict y by $\varphi(x, \mathcal{L})$. Now, suppose we are given a sequence of learnings sets $\{\mathcal{L}_k\}$, each consisting of N independent observations from the same underlying distribution as $\mathcal{L}$. Our mission is to use the $\{\mathcal{L}_k\}$ to get a better predictor than the single learning set predictor $\varphi(x, \mathcal{L})$. The approach is restricted to using the sequence of predictors $\{\varphi(x, \mathcal{L}_k)\}$.

In a classification setting, $\varphi(x, \mathcal{L})$ predicts a class $j \in \{1, \dots, J\}$, and aggregated predictions from $\varphi(x, \mathcal{L}_k)$ follow from either voting or averaging across the individual tree predictions. Let $N_j = nr\{k;\ \varphi(x, \mathcal{L}_k) = j\}$ and take $\varphi_A(x) = argmax_j N_j$, that is, the j for which N_j is maximum. Bagging used repeated bootstrap samples $\{\mathcal{L}^{(B)}\}$ from $\mathcal{L}$ to produce

$\{\varphi(x, \mathcal{L}^{(B)})\}$. Let the $\{\varphi(x, \mathcal{L}^{(B)})\}$ votes or averaging estimate $\varphi_B(x)$. The $\{\mathcal{L}^{(B)}\}$ form replicate data sets, each consisting of N cases, drawn at random but with replacement, from $\mathcal{L}$. This means that $\{\mathcal{L}^{(B)}\}$ are replicate data sets drawn from the bootstrap distribution approximating the distribution underlying $\mathcal{L}$.

A critical factor in whether bagging will improve accuracy is the stability of the procedure for constructing φ. Improvement will occur for unstable procedures where a small change in $\mathcal{L}$ can result in large changes in φ (e.g. classification tree model).

2.2.4 Variable importance measures

Although bagging typically improves accuracy over a single tree, the results of the bagged models can be more difficult to interpret since decision trees produce an easily interpretable diagram of results. Further, bagging complicates the interpretation in terms of which variables are most important to the model. Therefore, bagging algorithm increase prediction accuracy at the price of interpretability.

Although random forests algorithm is much more hard to interpret comparing to a single tree, the significance of each predictive variable could be described by mean decrease accuracy and mean decrease Gini in classification setting. In the case of random forests, we can add up the total amount that the Gini index or predictive accuracy is decreased by excluding a given predictor from the model, averaged over all trees. Predictors with greater mean decrease accuracy or mean decrease Gini has more importance in predicting the outcome.

2.2.5 Differences between random forests and bootstrap

Random forests algorithm provides a development over bagging by decorrelating the trees. Random forests first build a number of decision trees on bootstrapped training samples. And then, at each terminal node in a tree, random forests randomly select m predictors as divide candidate variables from all the p variables in the model. The split can even use only one of those m predictors. At each split, random forests will randomly select a fresh sample of m predictors to do the split, and usually the number of sampled factor is determined as $m \approx \sqrt{p}$, in other words, the number of predictors considered at each terminal node is approximately equal to the square root of the total number of predictors.

Thus, in building a random forest, each split in the tree considers a minority of the available predictors. This step is done to decorrelate the trees. To illustrate, consider the instance where the data set has one very strong predictor and numerous other moderately or weakly associated predictors. In this case, the most all or all of bagged trees will most likely include this strong predictor in the top split. Consequently, the bagged trees will look quite similar and produce highly correlated predictions. This result, of highly correlated predictions, will lead to only small reductions in variance as compared to averaging many uncorrelated predictions, and thus produce similar results between individual and bagged trees.

Random forests overcome this problem by forcing each split to consider only a subset of the predictors. Therefore, in the above scenario, on average $(p - m)/p$ of the splits will not consider the strong predictor, and so other predictors will have a greater chance of inclusion in the splits. We can think of this process as decorrelating the trees, thereby making the average of the resulting trees less variable and hence more reliable.

A key aspect of random forests is the choice of predictor subset size m. Using $m = p$ equates to bagging. Most commonly, random forests specify $m = \sqrt{p}$ to produce a reduction in both test error and out-of-bag (OOB) error over bagging. In contrast, using a smaller value of m in building a random forest will typically be helpful in the case of a large number of correlated predictors. As with bagging, random forests will overfit less with an increasing number of trees (B); therefore, B must be chosen sufficiently large to reduce the error rate.

2.2.6 Algorithm for random forests

Random forest is developed as followed:

1) Randomly draw k bootstrap samples from the original dataset with replacement. For each bootstrap, data not drawn into the sample define another sample called out-of-bag (OOB).
2) For each bootstrap sample, grow an unpruned classification tree. At each node of the tree, randomly pick m_{try} variables and pick the best split from those variables.
3) These k trees compose a random forest. Each tree gives a classification vote and estimated probability; the final decision can then use majority vote or average prediction.

Using OOB samples, random forest can construct a diverse predictor rank measure and compute the prediction accuracy of each predictor. In addition, randomization in predictors and sample selection also reduce the decision error. As described previously, $m_{try} = \sqrt{p}$ is a standard choice for m_{try}. [47] For the other two key parameters, k and node size (nodesize) we specified k=1000 and nodesize=3.

2.2.7 Validation and assessment of prognostic accuracy

Prognosis performance of all models was validated using the HALT study A data. Each observation was predicted Comparing to calculated false positive (F_P), true positive (T_P), false negative (F_N) and true negative (T_N).

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + F_P + T_N + F_N}$$

$$\text{Sensitivity} = \frac{T_P}{T_P + F_N}$$

$$\text{Specificity} = \frac{T_N}{F_P + T_N}$$

In addition, the ROC curve, which plots sensitivity (y-axis) by 1-specificity (x-axis), was presented to evaluate model performance. The area under the curve (AUC) measures discrimination capability of the proposed models. The AUC ranges from 0.5-1, with an AUC of 0.5 indicating no prediction ability, and 1 meaning ideal prediction. For medical applications, AUC greater than 0.8 are generally considered reflective of high accuracy.

3.0 RESULTS

3.1 DATA DESCRIPTION

Table 1-4 outlines basic characteristics of both CRISP and HALT datasets. For the outcome of interest, number (percent) of observations reaching CKD stage 3B are shown in the third and fourth column of each table. Table 1 shows that those who reach CKD stage 3B are older (with a mean age of 36 versus 29 years; p<0.001) with fewer African Americans (5% versus 14%; p=0.03). Birthweight was similar (6.8 versus 7.2 pounds; p=0.18) as was percent of females (58% versus 60%; p=0.75).

Table 1. Demographic data descriptive summaries for CRISP

Covariates	**N**	**Total**	**Did not reach CKD stage 3B 108(55.1%)**	**Reached CKD stage 3B 88(44.9%)**	**P-value**
Age in years*	196	32.15 (8.82)	29.21 (8.75)	35.74 (7.19)	<0.001
Birthweight in pounds*	128	7.02 (1.47)	7.21 (1.30)	6.81 (1.63)	0.18
Race**	196				0.03
• African American		19 (13.97)	15 (13.89)	4 (4.55)	
• Non-African American		117 (86.03)	93 (86.11)	84 (95.45)	
Gender**	196				0.75
• Male		80 (40.82)	43 (39.81)	37 (42.05)	
• Female		116 (59.18)	65 (60.19)	51 (57.95)	

* For continuous variables, statistics were showed in mean (standard deviation).
P-values were calculated by Wilcoxon rank sum test.
** For categorical variables, statistics were showed in count (column proportion).
P-values were calculated by chi-square test.

Table 2 shows frequencies of different categorizations of gene mutations. Those who reach CKD stage 3B have a higher percentage of PKD1 mutations (92% versus 72%; p=0.002 using three categories). Later analyses (in the multivariable analyses) group PKD2 and NMD as a singe

category. The other two categorizations, namely mutation strength group and the semi-continuous mutation strength score were not different between those who did or did not reach stage 3B (p=0.53 and 0.31, respectively). Later analyses (in the multivariable analyses) grouped PKD1 versus PKD2 and NMD with MSG to obtain additional variables consistent with guidance from the clinical investigators.

Table 2. Gene mutation data descriptive summaries for CRISP

Covariates	**Total (196)**	**Did not reach CKD stage 3b** 108(55.1%)	**Reached CKD stage 3b 88(44.9%)**	**P-value****
Gene type				0.002
• NMD***	12 (6.12)	9 (8.33)	3 (3.41)	
• PKD1	159 (81.12)	78 (72.22)	81 (92.05)	
• PKD2	25 (12.76)	21 (19.44)	4 (4.55)	
Mutation strength group (MSG)				0.53
• Truncating (MSG1)	126 (68.48)	71 (71.72)	55 (64.71)	
• Strong mutation (MSG2)	37 (20.11)	17 (17.17)	20 (23.53)	
• Weak mutation (MSG3)	21 (11.41)	11 (11.11)	10 (11.76)	
Semi-continuous mutation strength score (SCMSS)				0.31
• No mutation (SCMSS=0)	12 (6.12)	9 (8.33)	3(3.41)	
• Weak mutation (SCMSS=3)	17 (8.67)	7 (6.48)	10(11.63)	
• Strong mutation (SCMSS=4)	41 (20.92)	21 (19.44)	20(22.73)	
• Truncating (SCMSS=5)	126 (64.29)	71 (65.74)	55(62.50)	

* Statistics were showed in count (column proportion)
** p-values were calculated by chi-square test.
*** NMD stands for “no mutation detected”.

Table 3, similar to Table 1 for the CRISP data, described demographics stratified by the final outcome status. Results again show that those who reach CKD stage 3B are slightly but significantly older (with a mean age of 39 versus 37 years; p<0.001). Unlike CRISP, the percentages of African Americans and females were higher, although not significantly (4.4% versus 1.5%; p=0.08 and 58% versus 51%; p=0.23). Birthweight was not collected in HALT.

Table 3. Demographic data descriptive summaries for HALT

Covariates	Total (436)	Did not reach CKD stage 3b 344 (78.9%)	Reached CKD stage 3b 92 (21.1%)	P-value
Age*	37.26 (7.94)	36.77 (8.13)	39.10 (6.93)	0.02
Race**				0.08
• African American	9 (2.06)	5 (1.45)	4 (4.35)	
• Non-African American	427 (97.94)	339 (98.55)	88 (95.65)	
Gender**				0.23
• Male	227 (52.06)	174 (50.58)	53 (57.61)	
• Female	209 (47.94)	170 (49.42)	39 (42.39)	

* For continuous variables, statistics were showed in mean (standard deviation).
P-values were calculated by Wilcoxon rank sum test.
** For categorical variables, statistics were showed in count (column proportion).
P-values were calculated by chi-square test.

Table 4 shows frequencies of different categorizations of gene mutations for HALT. Those who reach CKD stage 3B have a higher percentage of PKD1 mutations (88% versus 71%; $p<0.001$ using three categories). The other two categorizations, namely mutation strength group and the semi-continuous mutation strength score were not different between those who did or did not reach stage 3B ($p=0.57$ and 0.77, respectively).

Table 4. Gene mutation data descriptive summaries for HALT

Covariates	Total (436)	Did not reach CKD stage 3b 344 (78.9%)	Reached CKD stage 3b 92 (21.1%)	P-value**
Gene type				<0.001
• NMD***	36 (8.26)	28 (8.14)	8 (8.70)	
• PKD1	324 (74.31)	243 (70.64)	81 (88.04)	
• PKD2	76 (17.43)	73 (21.22)	3 (3.26)	
Mutation strength group (MSG)				0.57
• Truncating (MSG1)	263 (65.75)	205 (64.87)	58 (69.05)	
• Strong mutation (MSG2)	80 (20.00)	63 (19.94)	17 (20.24)	
• Weak mutation (MSG3)	57 (14.25)	48 (15.19)	9 (10.71)	
Semi-continuous mutation strength (SCMSS)				0.77
• No mutation (SCMSS=0)	36 (8.26)	28 (8.14)	8 (8.70)	
• Weak mutation (SCMSS=3)	57 (13.07)	48 (13.95)	9 (9.78)	
• Strong mutation (SCMSS=4)	80 (18.35)	63 (18.31)	17 (18.48)	
• Truncating (SCMSS=5)	263 (60.32)	205 (59.59)	58 (63.04)	

* Statistics were showed in count (column proportion)
** p-values were calculated by chi-square test.
*** NMD stands for “no mutation detected”.

3.2 RESULTS FROM LOGISTIC REGRESSION

The likelihood ratio test for the overall model yielded a p-value of 0.0001 indicating that sex, race and birthweight together added significantly to current age in predicting CKD stage 3B. Further, the variables available at birth (gender, gene mutation, race), without birthweight also add significantly to a model that contains current age in predicting renal decline ($p < 0.001$). To determine if birthweight adds significantly to other factors available at birth, the nested logistic models (with and without birthweight) were fit in the subset with birthweight records; the LRT (using the optimal coding of gene mutation, which was PKD1 versus PKD2/NMD as described in the subsequent tables) shows that birthweight did not add significantly to the reduced model ($p = 0.159$). Given that more than one third of observations in CRISP were missing birthweight, and birthweight was not recorded in HALT, birthweight was dropped from further analyses.

Unadjusted logistic results for age, gender and race are shown in Table 5. Results for logistic models adjusted for different gene coding are shown in Table 6 and 7.

Table 5. Results of the unadjusted logistic model for demographics*

Covariates	Odds ratio (95% CI)	P-value**
Age	1.11 (1.06, 1.15)	<0.001
Male	1.10 (0.58, 2.06)	0.776
African American	0.24 (0.07, 0.82)	0.023

* LRT of all main effects versus the null for unadjusted logistic yielded p<0.001.
**p-values were calculated by LRT.

Table 6. Results of the logistic model for each gene coding adjusting for demographics

Covariates	Odds ratio (95% CI)	P-value*	10-fold cross validation AUC
Model 1		<0.001	0.7556
PKD2/NMD**			
PKD1	6.77 (2.55, 18.00)	<0.001	
Age	1.13 (1.08, 1.18)	<0.001	
Male	1.24 (0.64, 2.42)	0.5237	
African American	0.40 (0.11, 1.45)	0.1455	
Model 2		<0.001	0.7481
PKD2/NMD**			
PKD1-truncating	1.92 (0.91, 2.93)	<0.001	
PKD1-non-truncating	1.89 (0.81, 2.98)	0.001	
Age	0.12 (0.08, 0.16)	<0.001	
Male	0.20 (-0.47, 0.87)	0.550	
African American	-0.92 (-2.21, 0.37)	0.160	
Model 3		<0.001	0.7519
PKD2/NMD**			
PKD1_MSG1+2	7.04 (2.63, 18.87)	<0.001	
PKD1_MSG3	5.02 (1.28, 19.63)	0.020	
Age	1.13 (1.08, 1.18)	<0.001	
Male	1.24 (0.64, 2.41)	0.531	
African American	0.39 (0.11, 1.41)	0.150	
Model 4		<0.001	0.7450
SCMSS (=0) **			
Weak mutation (SCMSS=3)	0.73 (0.09, 5.82)	0.765	
Strong mutation (SCMSS=4)	0.50 (0.08, 3.24)	0.465	
Truncating (SCMSS=5)	0.52 (0.09, 3.06)	0.466	
PKD2			
PKD1	8.50 (2.51, 28.83)	0.001	
Age	1.13 (1.08, 1.18)	<0.001	
Male	1.26 (0.64, 2.48)	0.501	
African American	0.41 (0.11, 1.51)	0.182	

*p-values of models were calculated by LRT of all main effects versus the null.

*p-values of each main effect were obtained by LRT.
**The overall p-values for PKD1, PKD1-truncating, PKD1_MSG1+2 and the semi-continuous mutation strength score (SCMSS) in the above models were <0.0001, 0.0001, 0.0001 and 0.0010, respectively.

Table 7. 10-fold cross validation AUCs for logistic models adjusting for different gene mutation

	Model 1	**Model 2**	**Model 3**	**Model 4**
10-fold cross validation AUROC	0.7556	0.7481	0.7519	0.7450
95% confidence interval	(0.6644, 0.8038)	(0.6798, 0.8164)	(0.6839, 0.8200)	(0.6762, 0.8138)

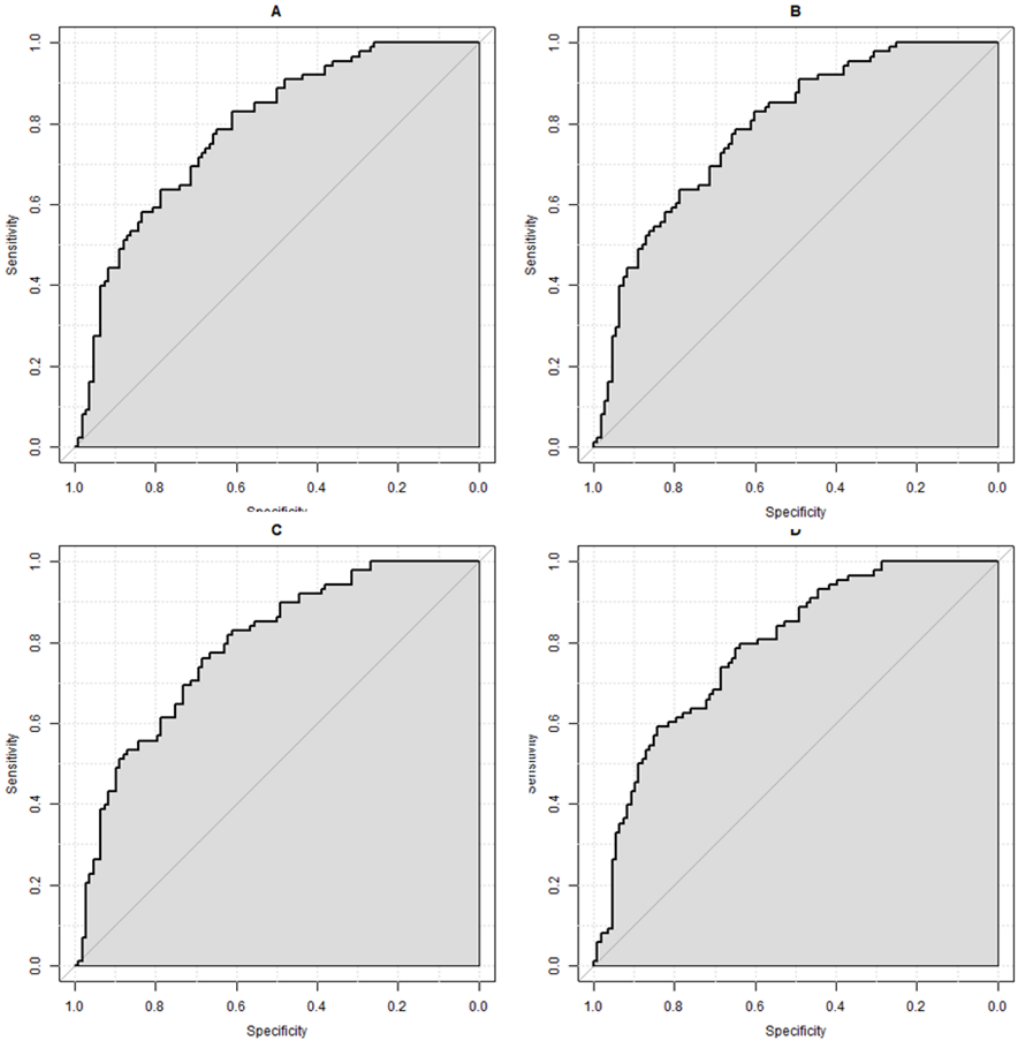

Figure 2. 10-fold cross validation ROC for logistic models adjusting for different gene mutation codings

(A) Model 1 **(B)** Model 2 **(C)** Model 3 **(D)** Model 4

The above results show that all four codings of gene mutation had statistically significant p-values, indicating that gene mutation information contributed significantly to prognosis of renal decline. 10-fold cross validation ROC curves for four models are shown in Figure 2. Among the four coding of gene mutation, the first model with PKD1 versus PKD2/NMD yielded the smallest BIC and largest 10-fold cross validated AUC, so this model was selected as the optimal coding of gene mutation. PKD1 mutations and older age were associated with higher odds of renal function decline. Both unadjusted and adjusted odds ratios and p-values and individual AUC for the selected best model are shown in Table 8.

Table 8. Coefficients of the best logistic model

Covariates	**Unadjusted Odds ratio (95% CI)**	**Adjusted odds ratio (95% CI)**	**P-value**	**Adjusted P-value**	**Individual AUC**
PKD2/NMD PKD1	4.45 (1.85, 10.72)	6.77 (2.55, 18.00)	<0.0001	<0.0001	0.5991
Age	1.10 (1.06, 1.15)	1.13 (1.08, 1.18)	<0.0001	<0.0001	0.7136
Male	1.09 (0.62, 1.94)	1.24 (0.64, 2.42)	0.7520	0.5237	0.5112
African American	0.30 (0.09, 0.92)	0.40 (0.11, 1.45)	0.0226	0.1455	0.5467

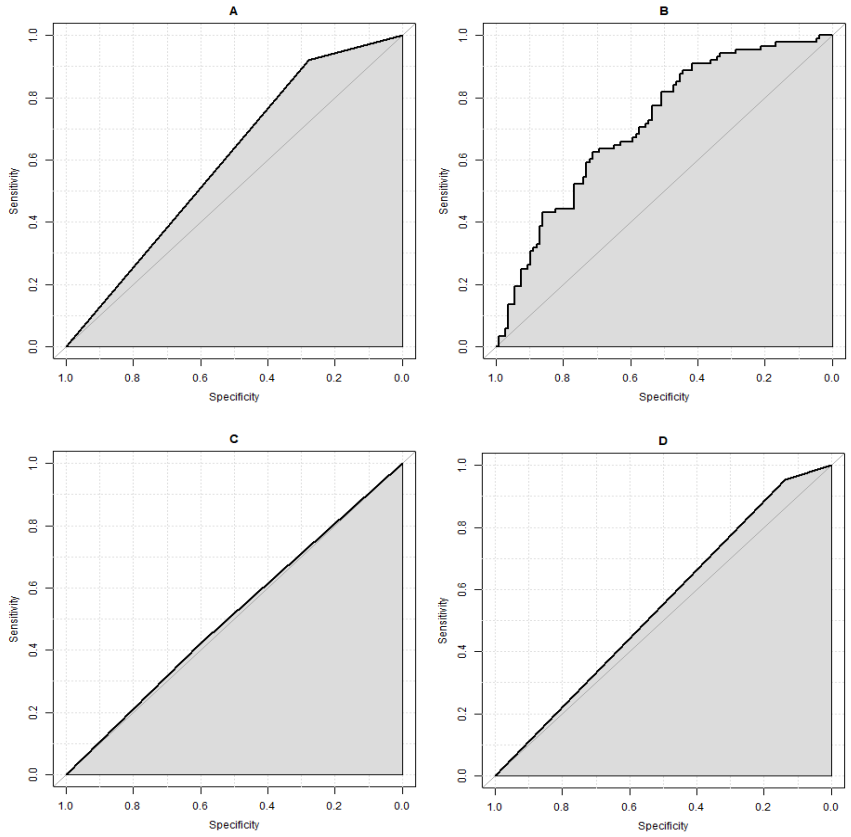

Figure 3. Individual ROC curves for each main effect in the best logistic model

(A) PKD1 vs. PKD2/NMD **(B)** age **(C)** gender **(D)** race

3.3 RESULTS FROM CLASSIFICATION TREES

The unpruned tree model is shown in Figure 4. Each number below each node represent the amount of observations with/without CKD stage 3B in corresponding node. At each splitting node, observations that agree with the splitting rule are shown in the left branch and others in the right branch. Results show that female patients over 35 years old, and with a mutation score above 0, are more likely to reach CKD stage 3B. Age accounted for the most splits.

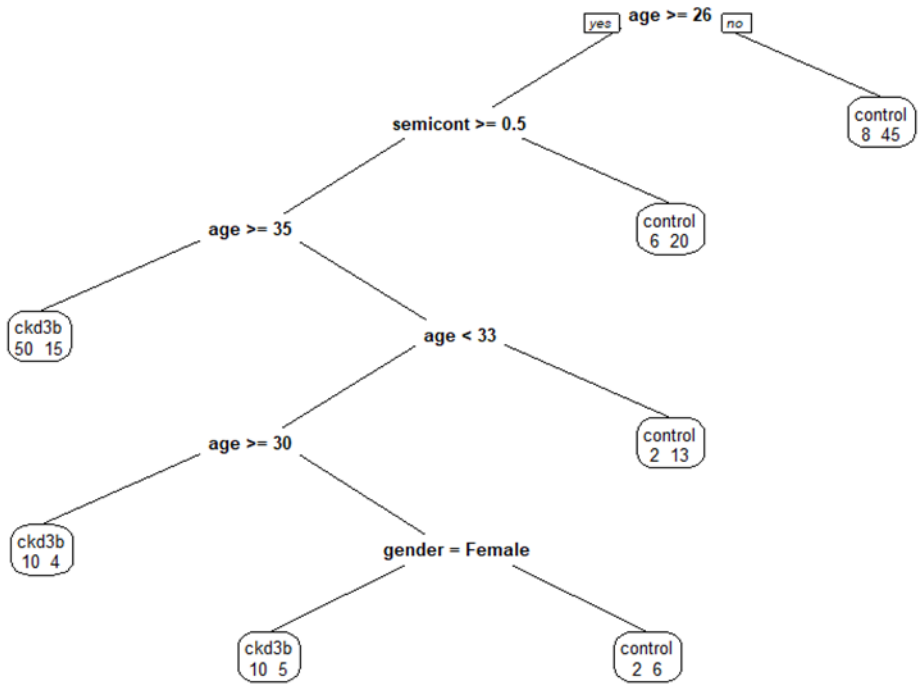

Figure 4. Tree model without pruning classifying CKD stage 3B

Figure 5 shows the relative misclassification error of tree model accompanying with each complexity parameter, resulting in an optimal value of 0.038. Figure 6 shows the pruned tree. After pruning, the classification tree only includes age and gene mutation, with a misclassification rate of 23.47%. As shown in Figure 7, classification tree model after pruning has a 10-fold cross validation AUC of 0.8008 (95% confidence interval: 0.7394, 0.8622), suggesting strong discrimination ability.

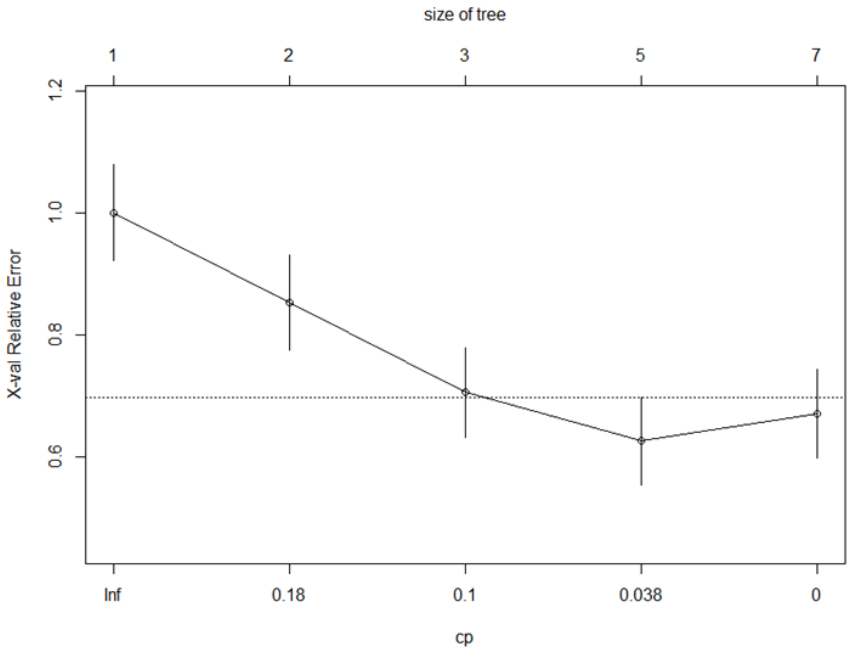

Figure 5. Cross-validation relative error for each complex parameter

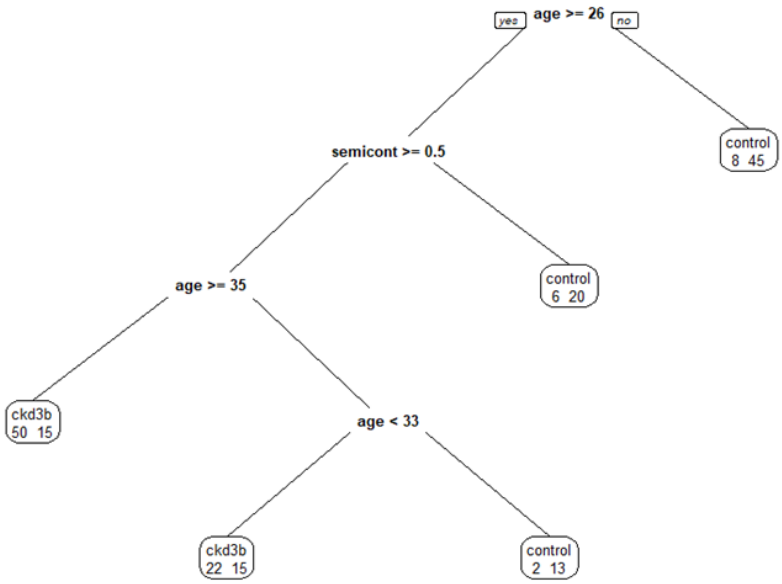

Figure 6. Pruned tree for predicting stage CKD 3B

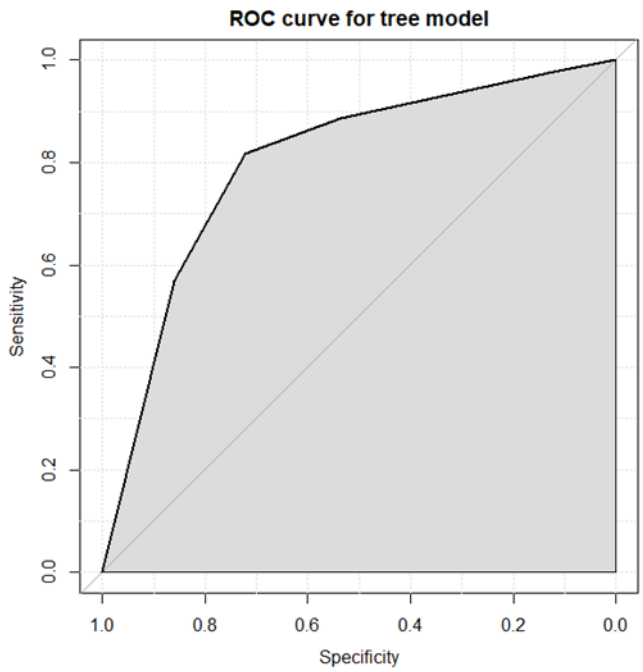

Figure 7. 10-fold cross validation ROC for pruned tree model

As a comparison, a pruned tree model with age alone (Figure 8) was fit, yielding an AUC of 0.7257 (95% CI of (0.6612, 0.7902)). The AUC and 95% CI for the tree and logistic models are listed in Table 9.

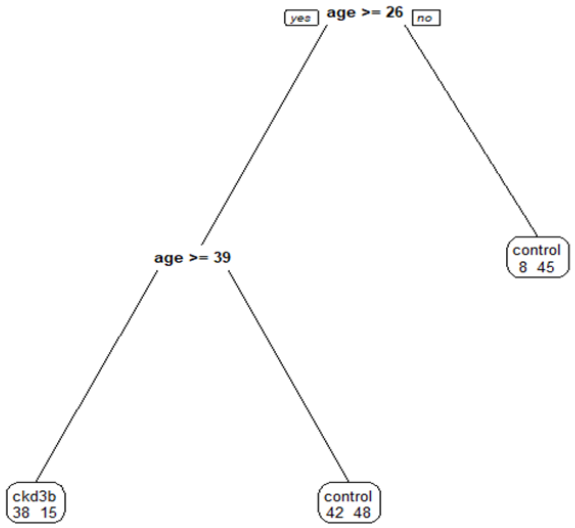

Figure 8. Pruned tree model with age alone

Table 9. Summary of AUC results of logistic regression and tree model

	Logistic model	**Tree model**	**Tree model with age alone**
10-fold cross validation AUROC (95% CI)	0.7556 (0.6644, 0.8038)	0.8008 (0.7394, 0.8622)	0.7257 (0.6612, 0.7902)

When comparing the tree model with all predictors to the tree model with age alone using only the training data, using DeLong's test for AUC, pruned tree model with all variables had a significantly better AUC (p=0.006778).

3.4 RESULTS FROM RANDOM FORESTS

Random forests, using 1000 bootstrap samples of size 196, yielded an out-of-bag misclassification rate 34.18% with voting. The variable importance measures are listed in Table 4 below.

Table 10. Importance of variables

	Mean decrease accuracy	**Mean decrease Gini**
age	30.93684	22.3869
SCMSS	-0.45672	2.402299
gene	10.10632	2.328235
PKD1 (2)	9.50293	1.80623
Truncating (3)	7.359186	2.7477
MSG (3)	9.642423	2.564204

Figure 9 shows the ROC curve for the random forest model, with an AUC of 0.9078, with 95% CI of (0.8689, 0.9467). Compared to logistic regression and the single classification tree, random forests yielded greater prognostic ability using the factors available at birth.

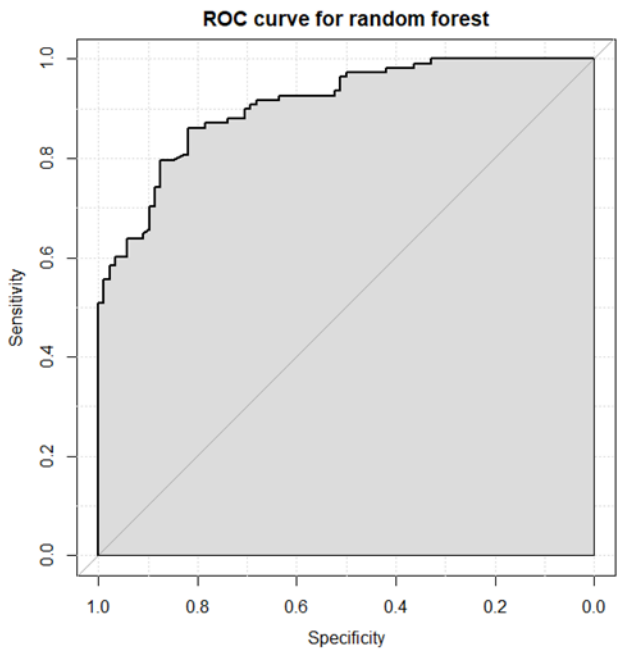

Figure 9. ROC for random forest model

3.5 MODEL VALIDATION AND DATA SPLITTING

Models fitted above based on the training dataset were then used to predict CKD stage 3B outcome using HALT dataset. AUC and 95% CIs for each model are shown in Table 11 with corresponding ROC curves shown in Figure 10.

Table 11. Summary of AUC results of validation

	Logistic model	**Tree model**	**Random forests**
AUC (95% CI)	0.5371 (0.4827, 0.5914)	0.5524 (0.5012, 0.6035)	0.6659 (0.5860, 0.7458)

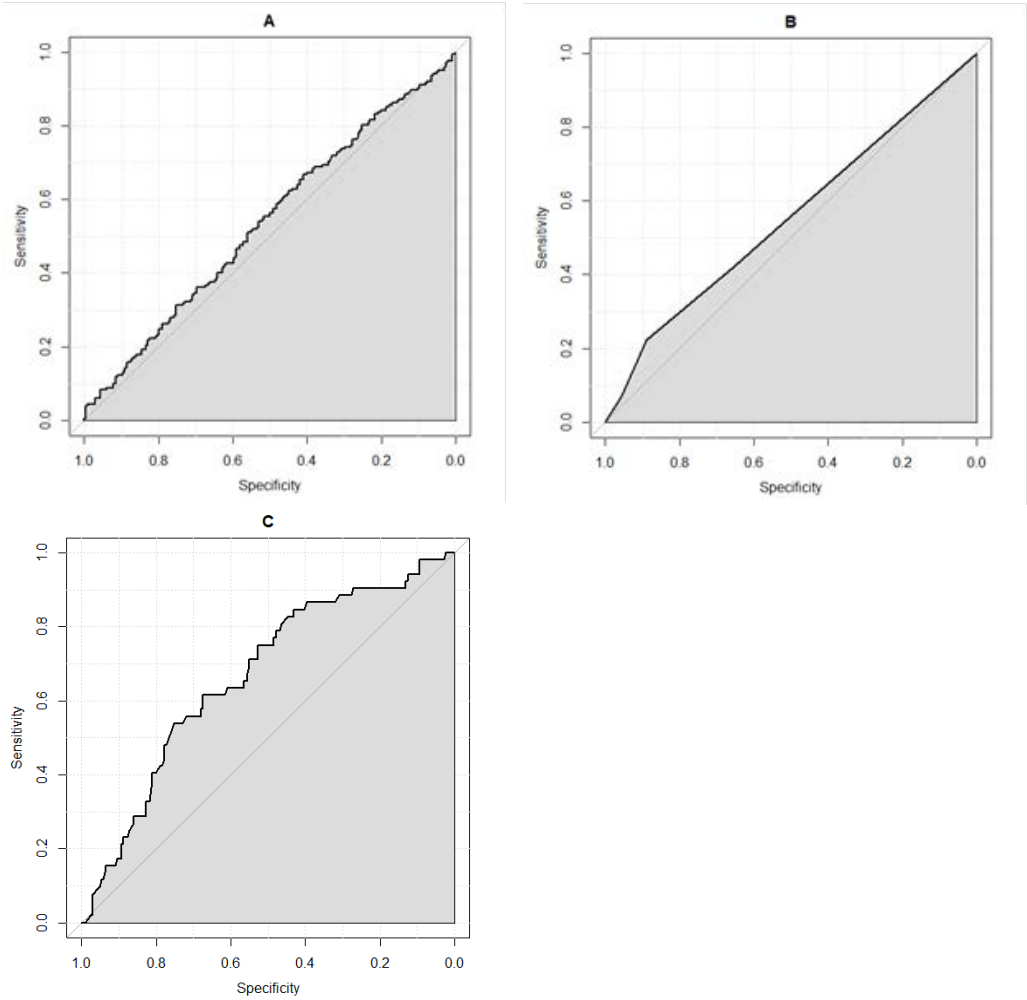

Figure 10. Validation ROC curves

(A) Logistic regression model **(B)** Pruned tree model **(C)** Random forests

The logistic and tree models established on CRISP dataset had similar, and quite limited ability to predict CKD outcomes on HALT study, while random forests had a higher, but still fairly low AUC. These validation results provided further evidence that random forests could improve prognostic ability of the factors available at birth. The relatively low AUC may be due to the shorter follow-up, and thus lower incidence of CKD stage 3B of HALT study A (~21%), compared to CRISP (~47%). Other inherent issues within the models like instability of tree model and overfitting of random forest may also lead to a low AUC in validation.

To further test if the reduction in accuracy with validation data is due to overfitting or systematic differences between CRISP and HALT data, additional analyses were performed within the CRISP data and compared to the HALT validation analysis, including the following analyses and statistics: 1) resubstitution accuracy was assessed using the entire CRISP data (n=194) for training and again for testing; 2) cross-validation accuracy was assessed using 10 randomly selected partitions of the data (with n=19-20 per partition); 10 separate models were then fit, each using 9 of the 10 partitions of data, with predictions calculated on the other (hold out) data; 3) data were randomly split only once into a train data (n=154) and test data (n=40); and 4) the AUCs for these models were then compared to HALT validation in Table 12.

Table 12. Summary table for model training and validation AUC

	Logistic regression	**Tree model**	**Random forests**
Resubstition	0.7861 (0.7231, 0.8491)	0.8193 (0.7424, 0.8818)	0.9328 (0.9007, 0.9649)
10-fold cross validation	0.7556 (0.6644, 0.8038)	0.8008 (0.7394, 0.8622)	0.9078 (0.8689, 0.9467)
CRISP training-test split	0.6793 (0.5015, 0.8571)	0.7340 (0.5586, 0.9094)	0.7609 (0.6028, 0.919)
HALT validation	0.5371 (0.4827, 0.5914)	0.5524 (0.5012, 0.6039)	0.6659 (0.5860, 0.7458)

Results in Table 12 show that 10-fold cross validation yields similar results to reubstitution , and those results are substantial better than results from the single training-test split, reflecting over-fitting in the resubstitution and 10-fold cross-validation results. Results also show an even greater degregation of classification accuracy between the single training-test split and the HALT

validation results, likely reflecting that the differences in length of follow-up are negatively affecting classification in HALT.

3.6 ADJUSTED COMPARISON BETWEEN THREE MODELS

Because logistic regression could not allow all the gene mutation information fitted in the model due to collinearity, this raised concern that the tree and random forest models benefit unfairly from the inclusion of multiple codings for gene mutation. To address this question, an adjusted logistic model using PKD1 versus PKD2 or NMD with high order of interactions (PKD1*age, PKD1*gender, PKD1*race and PKD1*age*race) was fitted. And we also fitted tree and random forest models using only the same gene mutation coding (PKD1 versus PKD2 or NMD) in adjusted logistic regression. Classification tree was pruned using minimal cost-complexity method. 10-fold cross validation was then performed, and models were compared using 10-fold cross validation AUC. Results are shown in Table 13.

From the results, adding high-order interactions increase AUC of logistic regression. Limiting gene information did not substantially affect prediction performance of tree model and random forests. This indicated that gene mutation codings were highly correlated and deleting redundant gene mutation information from the model would not harm the prognostic ability.

Table 13. AUC comparison between three models using PKD1 vs. PKD2/NMD gene mutation

Original models	**Logistic regression**	**Pruned tree model**	**Random forests**
10-fold CV AUC	0.7556	0.8008	0.9078
95% CI	(0.6644, 0.8038)	(0.7394, 0.8622)	(0.8689, 0.9467)
Adjusted models	**Logistic regression**	**Pruned tree model**	**Random forests**
10-fold CV AUC	0.7850	0.8027	0.9303
95% CI	(0.7219, 0.8482)	(0.7414, 0.8639)	(0.8978, 0.9629)

4.0 DISCUSSION

For classification CKD stage 3B, logistic regression suggests variables available at birth (gender, race, gene mutations and mutation strength) are significant predictors ($p<0.001$) to identify renal decline within a decade (or 15 years) for currently healthy PKD patients. Among variations of gene mutation information, PKD1 verses PKD2/NMD was the optimal coding with the smallest BIC of the model.

The optimal pruned tree model depended only on age and PKD1 mutations, with a cutoff value at age 35. Comparing to logistic regression, the tree model allowed for non-linear high order interactions. But tree model is unstable; the prediction of tree models might differ dramatically with even a slight change in the dataset.

Random forests greatly improved prognostic ability of the factors available at birth compared to other models. It also suggested high importance of age and gene mutations. Using validation with an independent dataset, random forest still yielded the greatest AUC. However, the much lower accuracy in the validation data may reflect overfitting and differences in length of follow-up. To address this limitation, CRISP will continue to follow HALT A participants over the next five years, and to achieve more at least 10 years of follow-up. For future extensions of this thesis, it would be of interest to consider other bias-correction approaches such as bootstrap or alternative cross-validation techniques.

There are limitations to the comparison of logistic regression and tree and random forest models, as the logistic model do not allow all the gene mutation information fitted in the same model due to collinearity, while tree and random forest models can benefit from the inclusion of multiple codings for gene mutation.

Results in this thesis pointed to the similar conclusions across different analyses, i.e. that variables available at birth could be used for modest gains in prognosis of renal decline with random forest models. In terms of public health significance, random forests could help estimate the probability of PKD patients reaching renal failure at given age, and thus inform prevention efforts. Although the AUC of about 0.6 with random forests may not seem very impressive, findings show improvement over what researchers had expected (which was no information from factors available at birth). Findings therefore make a significant contribution to the literature.

APPENDIX A: STATA CODES

```
/*  import data: 236 sample with gene info */

import delimited "C:\Users\tis42\OneDrive\thesis\data\variables at birth_236 sample.csv"


/* category data: demographic variables */

tab race ckd3b, col fre chi2

tab gender ckd3b, col fre chi2

sum age

sum age if ckd3b=="1"

sum age if ckd3b=="0"

ttest age, by(ckd3b) unequal

sum brwgt

sum brwgt if ckd3b=="1"

sum brwgt if ckd3b=="0"

ttest brwgt, by(ckd3b) unequal


/* gene variables */

tab gene ckd3b, col fre chi2

tab msg ckd3b, col fre chi2

tab scmss ckd3b, col fre chi2


/* convert string to numeric variables */
```

```
rename ckd3a ckd3a_c

rename ckd3b ckd3b_c

rename ckd4 ckd4_c

destring ckd3a_c, generate(ckd3a)

destring ckd3b_c, generate(ckd3b)

destring ckd4_c, generate(ckd4)


rename gender gender_c

gen gender=1 if gender_c=="Male"

replace gender=0 if gender_c=="Female"

label define male_label 1 "Male" 0 "Female"

label variable gender "male_label"

label values gender male_label


rename race race_c

gen race=1 if race_c=="African American"

replace race=0 if race_c=="Non African American"

label define race_label 1 "African American" 0 "Non African American"

label variable race "race_label"

label values race race_label


rename gene gene_c

gen gene=0 if gene_c=="NMD"
```

```
replace gene=1 if gene_c=="PKD1"

replace gene=2 if gene_c=="PKD2"

label define gene_label 0 "NMD" 1 "PKD1" 2 "PKD2"

label variable gene "gene_label"

label values gene "gene_label"


replace msg=0 if msg==.


/* fit full model with age with factors available at birth */

logit ckd3b age i.gender i.race i.gene i.msg

/* save current model */

estimates store full


/* Fit a logistic model with age only */

logit ckd3b age


/* Test with the likelihood ratio.*/

lrtest full


/* fit full model with age with factors available at birth. */

logit ckd3b age i.gender i.race i.gene i.msg brwgt

/* save current model */

estimates store full
```

```
/* Fit a logistic model with age only */
logit ckd3b age if brwgt!=.

/* Test with the likelihood ratio */
lrtest full

/* gene codings */
gene genecode1=1 if gene==1
replace genecode1=0 if gene!=1
label define genecode1 1 "pkd1" 0 "pkd2/nmd"
label values genecode1 genecode1

gen genecode2=0 if genecode1==0
replace genecode2=2 if genecode1==1 & trunc_grp=="Truncating"
replace genecode2=1 if genecode1==1 & trunc_grp=="Non Truncating"
label define genecode2 0 "pkd2/nmd" 1 "pkd1-non-truncating" 2 "pkd1-truncating"
label values genecode2 genecode2

rename genecompkd_msg genecompkd_msg_c
gene genecode3=1 if genecompkd_msg_c=="PKD1_MSG1+2"
replace genecode3=2 if genecompkd_msg_c=="PKD1_MSG3"
replace genecode3=0 if genecompkd_msg_c=="PKD2+NMD"
```

```
label define genecode3 1 "pkd1_msg1+2" 2 "pkd1_msg3" 0 "pkd2+nmd"
label values genecode3 genecode3

/* try different gene mutation codings */

logit ckd3b age i.gender i.race i.genecode1, or
estat ic
estimates store full
logit ckd3b age i.gender i.race, or
lrtest full

logit ckd3b age i.gender i.race i.genecode2 if genecode2!=.
estat ic
estimates store full
logit ckd3b age i.gender i.race if genecode2!=., or
lrtest full

logit ckd3b age i.gender i.race i.genecode3 if genecode3!=., or
estat ic
estimates store full
logit ckd3b age i.gender i.race if genecode3!=., or
lrtest full
```

```
logit ckd3b age i.gender i.race i.scmss i.gene, or
estat ic
estimates store full
logit ckd3b age i.gender i.race, or
lrtest full

logit ckd3b age i.gender i.race i.scmss i.gene 1.gene#i.scmss , or
estat ic

logit ckd3b age i.gender i.race i.genecode1, or
estat ic
lroc
estimates store full
logit ckd3b i.gender i.race i.genecode1, or
lrtest full
logit ckd3b age i.race i.genecode1, or
lrtest full
logit ckd3b age i.gender i.genecode1, or
lrtest full

logit ckd3b age
lrtest full
```

```
/* add brwgt */

logit ckd3b age i.gender i.race i.genecode1 brwgt

estimates store full

logit ckd3b age i.gender i.race i.genecode1 if brwgt!=.

lrtest full


/*improt halt data*/

use "C:\Users\tis42\OneDrive\thesis\halt\halt data for validation.dta"

/*data descriptive analysis*/

tab race ckd3b, col fre chi2

tab gender ckd3b, col fre chi2

sum age

sum age if ckd3b==1

sum age if ckd3b==0

ttest age, by(ckd3b) unequal


/* gene variables */

tab gene ckd3b, col fre chi2

tab msg ckd3b, col fre chi2

tab scmss ckd3b, col fre chi2
```

APPENDIX B: R CODES

```
# install.packages("rpart")
library(rpart)
# install.packages("rpart.plot")
library(rpart.plot)
# install.packages("randomForest")
library(randomForest)

# ====================
# data pre-processing
# ====================
getwd()
setwd("C:/Users/dell/onedrive/thesis/data")
sample<-read.csv("thesis data ckd3b_196 sample v2.csv")
# =======================
# data description
# =======================
setwd("C:/Users/dell/onedrive/thesis/data")
data<-read.csv("thesis data ckd3b_196 sample v2.csv")
data[!is.na(data$ckd3b_c) & data$ckd3b_c==1,]$ckd3b_c<-"ckd3b"
data[data$ckd3b_c==0,]$ckd3b_c<-"control"
```

```
# #split data into train and test
# set.seed(1021)
# a<-sample(1:196, 40, replace=F)
# #train data
# data<-sample[-a,]
# #test data
# test<-sample[a,]

do.classification <- function(train.set, test.set,
                              cl.name, verbose=F) {
  ## note: to plot ROC later, we want the raw probabilities,
  ## not binary decisions
  switch(cl.name,

         lr = { # logistic regression
           model = glm(ckd3b_c~age+gender+race+genecode1,
family=binomial(link="logit"), data=train.set)
           if (verbose) {
             print(summary(model))
           }
           prob = predict(model, newdata=test.set, type="response")
           #print(cbind(prob,as.character(test.set$y)))
           prob
```

```
},

dtree = { # decision tree
  model = rpart(ckd3b_c~age+gender+race+genecode1, data=train.set)
  if (verbose) {
    print(summary(model)) # detailed summary of splits
    printcp(model) # print the cross-validation results
    plotcp(model) # visualize the cross-validation results
    ## plot the tree
    plot(model, uniform=TRUE, main="Classification Tree")
    text(model, use.n=TRUE, all=TRUE, cex=.8)
  }
  prob = predict(model, newdata=test.set)

  if (0) { # here we use the default tree,
    ## you should evaluate different size of tree
    ## prune the tree
    pfit<- prune(model,
cp=model$cptable[which.min(model$cptable[,"xerror"]),"CP"])
    prob = predict(pfit, newdata=test.set)
    ## plot the pruned tree
    plot(pfit, uniform=TRUE,main="Pruned Classification Tree")
```

```
    text(pfit, use.n=TRUE, all=TRUE, cex=.8)
  }
  #print(cbind(prob,as.character(test.set$y)))

  prob = prob[,2]/rowSums(prob) # renormalize the prob.
  prob
},

dtreeprune = {
  model = rpart(ckd3b_c~age+gender+race+genecode1, data=train.set)
  if (verbose) {
    print(summary(model)) # detailed summary of splits
    printcp(model) # print the cross-validation results
    plotcp(model) # visualize the cross-validation results
    ## plot the tree
    plot(model, uniform=TRUE, main="Classification Tree")
    text(model, use.n=TRUE, all=TRUE, cex=.8)
  }
  prob = predict(model, newdata=test.set)

  if (1) { # here we prune the tree
    ## prune the tree
```

```
      pfit<- prune(model,
cp=model$cptable[which.min(model$cptable[,"xerror"]),"CP"])
      prob = predict(pfit, newdata=test.set)
      ## plot the pruned tree
      plot(pfit, uniform=TRUE,main="Pruned Classification Tree")
      text(pfit, use.n=TRUE, all=TRUE, cex=.8)
    }
    #print(cbind(prob,as.character(test.set$y)))
    prob = prob[,2]/rowSums(prob) # renormalize the prob.
    prob
  },


  rf = { # random forests
    model = randomForest(ckd3b_c~age+gender+race+genecode1, data=train.set)
    if (verbose) {
      print(summary(model)) # detailed summary of random forests


    }
    prob = predict(model, newdata=test.set)
    prob
  }
)
}
```

```
#10-fold cross validation
k.fold.cv <- function(dataset, cl.name, k.fold=10, prob.cutoff=0.5, get.performance=F) {
  ## default: 10-fold CV, cut-off 0.5
  n.obs <- nrow(dataset) # no. of observations
  s = sample(n.obs)
  errors = dim(k.fold)
  precisions = dim(k.fold)
  recalls = dim(k.fold)
  fscores = dim(k.fold)
  accuracies = dim(k.fold)
  probs = NULL
  actuals = NULL
  for (k in 1:k.fold) {
    test.idx = which(s %% k.fold == (k-1) ) # use modular operator
    train.set = dataset[-test.idx,]
    test.set = dataset[test.idx,]
    cat(k.fold,'-fold CV run',k,cl.name,':',
        '#training:',nrow(train.set),
        '#testing',nrow(test.set),'\n')
    prob = do.classification(train.set, test.set, cl.name)
    predicted = as.numeric(prob > prob.cutoff)
    actual = test.set$y
```

```
  confusion.matrix = table(actual,factor(predicted,levels=c(0,1)))
  confusion.matrix
  error = (confusion.matrix[1,2]+confusion.matrix[2,1]) / nrow(test.set)
  errors[k] = error
  cat('\t\terror=',error,'\n')
  precision = (confusion.matrix[1,1]/(confusion.matrix[1,1]+confusion.matrix[2,1]))
  precisions[k] = precision
  print(confusion.matrix)
  recall =(confusion.matrix[1,1]/(confusion.matrix[1,1]+confusion.matrix[1,2]))
  recalls[k] = recall
  probs = c(probs,prob)
  actuals = c(actuals,actual)
  ## you may compute other measures and store them in arrays
}

avg.error = mean(errors)
cat('avg error=',avg.error,'\n')
avg.accuracy = 1 - avg.error
cat('avg Accuracy=',avg.accuracy,'\n')
avg.precision = mean(precisions)
cat('avg Precision=',avg.precision,'\n')
avg.recall = mean(recalls)
cat('avg recall=',avg.recall,'\n')
```

```
  avg.fscore = mean(fscores)
  cat('avg fscore=',avg.fscore,'\n')

  ## plot ROC
  result = data.frame(probs,actuals)
  pred = prediction(result$probs,result$actuals)
  perf = performance(pred, "tpr","fpr")
  auc = performance(pred,"auc")
  auc <- as.numeric(auc@y.values)
  cat(k.fold,'-fold AUC:',auc,'\n')
  plot(perf)

#logit
logit<-glm(ckd3b_c ~ age + gender + race + genecode1,
           family=binomial(link='logit'),data=data)
#prediction
pred.logit<-predict(logit,newdata=data)
outcome<-cbind(data$pkdid,
               data$ckd3b_c,
               pred.logit)
colnames(outcome)<-c("id","ckd3b","pruned")
outcome<-as.data.frame(outcome)
```

```
roc1<-roc(ckd3b~pruned,outcome,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for logistic model")
# Area under the curve: 0.7861
# 95% CI: 0.7231-0.8491 (DeLong)
table(pred.logit, data$ckd3b_c)

#within test
pred.logit<-predict(logit,newdata=test)
outcome<-cbind(test$pkdid,
            test$ckd3b_c,
            pred.logit)
colnames(outcome)<-c("id","ckd3b","pruned")
outcome<-as.data.frame(outcome)

roc1<-roc(ckd3b~pruned,outcome,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for logistic model")

# high level interaction test
logit<-glm(ckd3b_c ~ age + gender + race + genecode1
         + age*genecode1 + gender*genecode1 + race*genecode1
         +age*race*genecode1,
         family=binomial(link='logit'),data=data)
summary(logit)
```

```
pred.logit<-predict(logit,newdata=data)
outcome<-cbind(data$pkdid,
               data$ckd3b_c,
               pred.logit)
colnames(outcome)<-c("id","ckd3b","pruned")
outcome<-as.data.frame(outcome)

roc1<-roc(ckd3b~pruned,outcome,ci=T,
          auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for logistic model")
# Area under the curve: 0.785
# 95% CI: 0.7219-0.8482 (DeLong)

#roc curves for each gene codings
setwd("C:/Users/dell/onedrive/thesis/gene code roc")
#coding 1
logit<-glm(ckd3b_c ~ age + gender + race + genecode1,
           family=binomial(link='logit'),data=data)
pred.logit<-predict(logit,newdata=data)
outcome<-cbind(data$pkdid,
               data$ckd3b_c,
               pred.logit)
colnames(outcome)<-c("id","ckd3b","pruned")
outcome<-as.data.frame(outcome)
```

```
png("gene code1.png")
roc1<-roc(ckd3b~pruned,outcome,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="A")
dev.off()
#coding 2
logit<-glm(ckd3b_c ~ age + gender + race + genecode2,
         family=binomial(link='logit'),data=data)
pred.logit<-predict(logit,newdata=data)
outcome<-cbind(data$pkdid,
            data$ckd3b_c,
            pred.logit)
colnames(outcome)<-c("id","ckd3b","pruned")
outcome<-as.data.frame(outcome)
roc1<-roc(ckd3b~pruned,outcome,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="B")
png("gene code2.png")
roc1<-roc(ckd3b~pruned,outcome,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="B")
dev.off()

#coding 3
logit<-glm(ckd3b_c ~ age + gender + race + genecode3,
         family=binomial(link='logit'),data=data)
```

```
pred.logit<-predict(logit,newdata=data)
outcome<-cbind(data$pkdid,
               data$ckd3b_c,
               pred.logit)
colnames(outcome)<-c("id","ckd3b","pruned")
outcome<-as.data.frame(outcome)
roc1<-roc(ckd3b~pruned,outcome,ci=T,
          auc.polygon=TRUE, grid=TRUE, plot=T, main="C")
png("gene code3.png")
roc1<-roc(ckd3b~pruned,outcome,ci=T,
          auc.polygon=TRUE, grid=TRUE, plot=T, main="C")
dev.off()


#coding 2
logit<-glm(ckd3b_c ~ age + gender + race + gene + scmss,
           family=binomial(link='logit'),data=data)
pred.logit<-predict(logit,newdata=data)
outcome<-cbind(data$pkdid,
               data$ckd3b_c,
               pred.logit)
colnames(outcome)<-c("id","ckd3b","pruned")
outcome<-as.data.frame(outcome)
roc1<-roc(ckd3b~pruned,outcome,ci=T,
```

```
        auc.polygon=TRUE, grid=TRUE, plot=T, main="D")
png("gene code4.png")
roc1<-roc(ckd3b~pruned,outcome,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="D")
dev.off()

#============
#tree model
fit.ckd3b<-rpart(formula = ckd3b_c ~ age + gender + race +
        semicontinuousmspkd1 + semicontinuousmspkd2 +
        scmss + msg + trunc_grp + gene +
        genecode1 + genecode2 + genecode3,
        data = data, method = "class",
        control=rpart.control(cp=0, xval=195))
prp(fit.ckd3b, faclen = 0, cex=0.8, extra = 1)
plotcp(fit.ckd3b)

fit.ckd3b.1<-prune(fit.ckd3b, cp=0.038)
prp(fit.ckd3b.1, faclen = 0, cex=0.8, extra = 1)
printcp(fit.ckd3b.1)

#prediction
pred.tree.ckd3b<-predict(fit.ckd3b.1,
```

```
                    type="prob",
                    newdata=data)
outcome.ckd3b<-cbind(data$pkdid,
                 data$ckd3b_c,
                 pred.tree.ckd3b[,2])
colnames(outcome.ckd3b)<-c("id","ckd3b","pruned")
outcome.ckd3b<-as.data.frame(outcome.ckd3b)

roc1<-roc(ckd3b~pruned,outcome.ckd3b,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for tree model")

#test
pred.tree.ckd3b<-predict(fit.ckd3b,
                    type="prob",
                    newdata=test)
outcome.ckd3b<-cbind(test$pkdid,
                 test$ckd3b_c,
                 pred.tree.ckd3b[,2])
colnames(outcome.ckd3b)<-c("id","ckd3b","pruned")
outcome.ckd3b<-as.data.frame(outcome.ckd3b)

roc1<-roc(ckd3b~pruned,outcome.ckd3b,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for tree model")
```

```
# test gene information
fit.ckd3b<-rpart(formula = ckd3b_c ~ age + gender + race + genecode1,
                data = data, method = "class",
                control=rpart.control(cp=0, xval=195))
plotcp(fit.ckd3b)
fit.ckd3b.1<-prune(fit.ckd3b, cp=0.038)
prp(fit.ckd3b.1, faclen = 0, cex=0.8, extra = 1)
printcp(fit.ckd3b.1)

#prediction
pred.tree.ckd3b<-predict(fit.ckd3b.1,
                     type="prob",
                     newdata=data)
outcome.ckd3b<-cbind(data$pkdid,
                  data$ckd3b_c,
                  pred.tree.ckd3b[,2])
colnames(outcome.ckd3b)<-c("id","ckd3b","pruned")
outcome.ckd3b<-as.data.frame(outcome.ckd3b)

roc1<-roc(ckd3b~pruned,outcome.ckd3b,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for tree model")
# Area under the curve: 0.8027
```

```
# 95% CI: 0.7414-0.8639 (DeLong)
fit.ckd3b.age<-rpart(formula = ckd3b_c ~ age,
                data = data, method = "class",
                control=rpart.control(cp=0, xval=195))
prp(fit.ckd3b.age, faclen = 0, cex=0.8, extra = 1)
plotcp(fit.ckd3b.age)

fit.ckd3b.age.1<-prune(fit.ckd3b.age, cp=0.052)
prp(fit.ckd3b.age.1, faclen = 0, cex=0.8, extra = 1)
printcp(fit.ckd3b.age.1)

#prediction
pred.tree.ckd3b.age<-predict(fit.ckd3b.age.1,
                     type="prob",
                     newdata=data)
outcome.ckd3b.age<-cbind(data$pkdid,
                   data$ckd3b_c,
                   pred.tree.ckd3b.age[,1])
colnames(outcome.ckd3b.age)<-c("id","ckd3b","pruned")
outcome.ckd3b.age<-as.data.frame(outcome.ckd3b.age)

roc.age<-roc(ckd3b~pruned,outcome.ckd3b.age,ci=T,
```

```
auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for tree model with age only")
##################################################################
# random forest

set.seed(1021)
data$ckd3b_c<-factor(data$ckd3b_c)
set.seed(1021)
rf.3b<-randomForest(ckd3b_c ~ age + gender + race +
                    scmss + gene +
                    genecode1 + genecode2 + genecode3,
                   data = data, na.action = na.omit, importance=T, ntree=1000)
importance(rf.3b, type=1)
importance(rf.3b, type = 2)

plot(rf.3b)
#prediction
p2 <- predict(rf.3b, type="prob",data)
outcome.rf<-cbind(data$pkdid,
                  data$ckd3b_c,
                  p2[,1])
colnames(outcome.rf)<-c("id","ckd3b","predict")
outcome.rf<-as.data.frame(outcome.rf)
```

```
roc.rf<-roc(ckd3b~predict,outcome.rf,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for random forest")
# Area under the curve: 0.9078
# 95% CI: 0.8689-0.9467 (DeLong)

#within test
p2 <- predict(rf.3b, type="prob",test)
outcome.rf<-cbind(test$pkdid,
                test$ckd3b_c,
                p2[,1])
colnames(outcome.rf)<-c("id","ckd3b","predict")
outcome.rf<-as.data.frame(outcome.rf)

roc.rf<-roc(ckd3b~predict,outcome.rf,ci=T,
          auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for random forest")

#age only
set.seed(1)
rf.3b.age<-randomForest(ckd3b_c ~ age,
                  data = data, na.action = na.omit, importance=T, ntree=1000)
p.age <- predict(rf.3b.age,data)
head(p.age)
```

```
confusionMatrix(p.age, data$ckd3b_c)


outcome.rf.age<-cbind(data$pkdid,
                data$ckd3b_c,
                p.age)
colnames(outcome.rf.age)<-c("id","ckd3b","predict")
outcome.rf.age<-as.data.frame(outcome.rf.age)


roc.rf.age<-roc(ckd3b~predict,outcome.rf.age,ci=T,
          auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for random forest
with age only")
# Area under the curve: 0.9916
# 95% CI: 0.9842-0.999 (DeLong)
#-----------------------------------------


# test interaction of gene mutation
set.seed(1021)
data$ckd3b_c<-factor(data$ckd3b_c)
rf.3b<-randomForest(ckd3b_c ~ age + gender + race + genecode1,
                  data = data, na.action = na.omit, importance=T, ntree=1000)
p2 <- predict(rf.3b, type="prob",data)
outcome.rf<-cbind(data$pkdid,
                data$ckd3b_c,
```

```
                p2[,1])
colnames(outcome.rf)<-c("id","ckd3b","predict")
outcome.rf<-as.data.frame(outcome.rf)
roc.rf<-roc(ckd3b~predict,outcome.rf,ci=T,
          auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for random forest")
# Area under the curve: 0.9303
# 95% CI: 0.8978-0.9629 (DeLong)


# validation using HALT data
# read in HALT data
setwd("C:/Users/tis42/onedrive/thesis/halt")
data<-read.csv("halt data for validation.rdata")


#----------------------
# fit logit model
logit<-glm(ckd3b_c ~ age + gender + race + genecode1,
          family=binomial(link='logit'),data=data)
#predict using HALT
pred.logit<-predict(logit,newdata=halt.data)
outcome<-cbind(halt.data$ckd3b,
             pred.logit)
colnames(outcome)<-c("ckd3b","pruned")
outcome<-as.data.frame(outcome)
```

```
roc1<-roc(ckd3b~pruned,outcome,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="ROC curve for logistic model")
# Area under the curve: 0.6584
# 95% CI: 0.5975-0.7194 (DeLong)


#-----------------------
# fit classification tree model
fit.ckd3b<-rpart(formula = ckd3b_c ~ age + gender + race +
                 semicontinuousmspkd1 + semicontinuousmspkd2 +
                 scmss + msg + trunc_grp + gene +
                 genecode1 + genecode2 + genecode3,
               data = data, method = "class",
               control=rpart.control(cp=0, xval=195))
prp(fit.ckd3b, faclen = 0, cex=0.8, extra = 1)
plotcp(fit.ckd3b)
fit.ckd3b.1<-prune(fit.ckd3b, cp=0.038)
prp(fit.ckd3b.1, faclen = 0, cex=0.8, extra = 1)
#predict using HALT
pred.tree.ckd3b<-predict(fit.ckd3b.1,
                     type="prob",
                     newdata=halt.data)
outcome.ckd3b<-cbind(halt.data$ckd3b,
                   pred.tree.ckd3b)
```

```
colnames(outcome.ckd3b)<-c("ckd3b","pruned")
outcome.ckd3b<-as.data.frame(outcome.ckd3b)
roc1<-roc(ckd3b~pruned,outcome.ckd3b,ci=T,
        auc.polygon=TRUE, grid=TRUE, plot=T, main="B")
# Area under the curve: 0.6332
# 95% CI: 0.5733-0.6931 (DeLong)


#-----------------------
# fit random forest model
data$ckd3b_c<-factor(data$ckd3b_c)
set.seed(1021)
rf.3b<-randomForest(ckd3b_c ~ age + gender + race + gene + scmss +
                  genecode1 + genecode2 + genecode3,
                 data = data, na.action = na.omit, importance=T, ntree=1000)
# predict using CRISP
p2 <- predict(rf.3b, type="prob",data)
outcome.rf<-cbind(data$ckd3b_c,
                p2[,1])
colnames(outcome.rf)<-c("ckd3b","predict")
outcome.rf<-as.data.frame(outcome.rf)
roc.rf<-roc(ckd3b~predict,outcome.rf,ci=T,
          auc.polygon=TRUE, grid=TRUE, plot=T, main="A")
```

```
# predict using HALT
p2 <- predict(rf.3b, type="prob",halt.data)
outcome.rf<-cbind(halt.data$ckd3b_c,
                  p2[,1])
colnames(outcome.rf)<-c("ckd3b","predict")
outcome.rf<-as.data.frame(outcome.rf)
roc.rf<-roc(ckd3b~predict,outcome.rf,ci=T,
            auc.polygon=TRUE, grid=TRUE, plot=T, main="C")
# Area under the curve: 0.6659
# 95% CI: 0.586-0.7458 (DeLong)
```

BIBLIOGRAPHY

[1] Gabow PA: Autosomal dominant polycystic kidney disease. N Engl J Med 329: 332–342, 1993

[2] Torres VE, Harris PC, Pirson Y: Autosomal dominant polycystic kidney disease. Lancet 369: 1287–1301, 2007

[3] Grantham JJ: The etiology, pathogenesis, and treatment of autosomal dominant polycystic kidney disease: recent advances. American journal of kidney diseases 28: 788–803, 1996.

[4] Grantham JJ: Autosomal dominant polycystic kidney disease. N Engl J Med 359: 1477-1485, 2008.

[5] Torres VE, Grantham JJ: Cystic diseases of the kidney. The kidney 6: 1428-1462, 2008.

[6] Simms RJ: Autosomal dominant polycystic kidney disease. Bmj 352: i679.

[7] Bajwa ZH, Gupta S, Warfield CA, Steinman TI: Pain management in polycystic kidney disease. Kidney Int 60: 1631-44, 2001.

[8] Bahwa ZH, Sial KA, Malik AB, Steinman TI: Pain patterns in patients with polycystic kidney disease. Kidney Int 66: 1561-9, 2004

[9] Ecder T, Schrier RW: Hypertension in autosomal-dominant polycystic kidney disease: early occurrence and unique aspects. J Am Soc Nephrol 12: 194-200, 2001.

[10] Bae KT, Zhu F, Chapman AB, et al. Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP). Magnetic resonance imaging evaluation of hepatic cysts in early autosomal-dominant polycystic kidney disease: the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease cohort. Clin J Am Soc Nephrol 1: 64-9, 2006;

[11] Peltola, P., Lumiaho, A., Miettinen, R. et al. J Mol Med (2005) 83: 638.

[12] Gall, Emilie Cornec-Le et al. “PKD2-Related Autosomal Dominant Polycystic Kidney Disease: Prevalence, Clinical Presentation, Mutation Spectrum, and Prognosis.” Am J Kidney Dis (2017) 70: 476–485.

[13] Heyer, C. M., et al. "Predicted mutation strength of nontruncating PKD1 mutations aids genotype-phenotype correlations in autosomal dominant polycystic kidney disease." Journal of the American Society of Nephrology 27.9 (2016): 2872-2884.

[14] Torres, Vicente E., et al. "Magnetic resonance measurements of renal blood flow and disease progression in autosomal dominant polycystic kidney disease." Clinical Journal of the American Society of Nephrology 2.1 (2007): 112-120.

[15] Chapman, Arlene B., et al. "Kidney volume and functional outcomes in autosomal dominant polycystic kidney disease." Clinical Journal of the American Society of Nephrology 7.3 (2012): 479-486.

[16] Ta, Michelle HT, David CH Harris, and Gopala K. Rangan. "Role of interstitial inflammation in the pathogenesis of polycystic kidney disease." Nephrology 18.5 (2013): 317-330.

[17] Tao, Yunxia, et al. "Rapamycin markedly slows disease progression in a rat model of polycystic kidney disease." Journal of the American Society of Nephrology 16.1 (2005): 46-51.

[18] Vanholder R, Massy Z, Argiles A, Spasovski G, Verbeke FR, Lameire N: Chronic kidney disease as cause of cardiovascular morbidity and mortality. Nephrology Dialysis Transplantation 20(6): 1048-56, 2005.

[19] Coresh J, Selvin E, Stevens LA, Manzi J, Kusek JW, Eggers P, Van Lente F, Levey AS: Prevalence of chronic kidney disease in the United States. Jama 17: 2038-47, 2007.

[20] Center for Chronic Disease Prevention. National Chronic Kidney Disease Fact Sheet, 2017.

[21] Stevens PE, Levin A. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. Annals of internal medicine. 2013 Jun 4;158(11):825-30.

[22] Jiun-Ruey Hu, Josef Coresh; The public health dimension of chronic kidney disease: what we have learnt over the past decade. Nephrol Dial Transplant 2017; 32.

[23] Kirsztajn, Gianna Mastroianni, Jose HR Suassuna, and Marcus G. Bastos. "Dividing stage 3 of chronic kidney disease (CKD): 3A and 3B." Kidney international 76.4 (2009): 462-463.

[24] Levey, Andrew S., et al. A new equation to estimate glomerular filtration rate. Annals of internal medicine 150.9 (2009): 604-612.

[25] Torres, Vicente E., et al. "Tolvaptan in patients with autosomal dominant polycystic kidney disease." New England Journal of Medicine 367.25 (2012): 2407-2418.

[26] Torres, Vicente E., et al. "Angiotensin blockade in late autosomal dominant polycystic kidney disease." New England Journal of Medicine 371.24 (2014): 2267-2276.

[27] Greene, T., Teng, C.C., Inker, L.A., Redd, A., Ying, J., Woodward, M., Coresh, J. and Levey, A.S., 2014. Utility and validity of estimated GFR–based surrogate time-to-event end points in CKD: a simulation study. American Journal of Kidney Diseases, 64(6), pp.867-879.

[28] Grantham JJ, Torres VE, Chapman AB, Guay-Woodford LM, Bae KT, King Jr BF, Wetzel LH, Baumgarten DA, Kenney PJ, Harris PC, Klahr S. Volume progression in polycystic kidney disease. New England Journal of Medicine. 2006 May 18;354(20):2122-30.

[29] Chapman AB, Bost JE, Torres VE, Guay-Woodford L, Bae KT, Landsittel D, Li J, King BF, Martin D, Wetzel LH, Lockhart ME. Kidney volume and functional outcomes in autosomal dominant polycystic kidney disease. Clinical Journal of the American Society of Nephrology. 2012 Mar 1;7(3):479-86.

[30] Zeleny Michael R. Assessing Neural Network Prediction in Kidney Disease Data. Master's Thesis, University of Pittsburgh. 2016.

[31] Gao Xiaotian: Analysis of kidney volume and functional outcomes using survival and classification tree models. Master's Thesis, University of Pittsburgh. 2015.

[32] Cabrera, Alberto F. "Logistic regression analysis in higher education: An applied perspective." Higher education: Handbook of theory and research 10 (1994): 225-256.

[33] Breiman, L., et al. "Regression Trees. Wadsworth Int." Group (1984).

[34] Podgorelec, Vili, et al. "Decision trees: an overview and their use in medicine." Journal of medical systems 26.5 (2002): 445-463.

[35] Breiman, Leo. Classification and regression trees. Routledge, 2017.

[36] Dwyer, Kenneth, and Robert Holte. "Decision tree instability and active learning." European Conference on Machine Learning. Springer, Berlin, Heidelberg, 2007.

[37] Breiman, Leo. "Bagging predictors." Machine learning 24.2 (1996): 123-140.

[38] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[39] Chapman, Arlene B., et al. "The HALT polycystic kidney disease trials: design and implementation." Clinical Journal of the American Society of Nephrology 5.1 (2010): 102-109.

[40] Torres, Vicente E., et al. "Analysis of baseline parameters in the HALT polycystic kidney disease trials." Kidney international 81.6 (2012): 577-585.

[41] Li, S., Chen, S.C., Shlipak, M., Bakris, G., McCullough, P.A., Sowers, J., Stevens, L., Jurkovitz, C., McFarlane, S., Norris, K. and Vassalotti, J., 2008. Low birth weight is associated with chronic kidney disease only in men. Kidney international, 73(5), pp.637-642.

[42] Torres, V.E., Grantham, J.J., Chapman, A.B., Mrug, M., Bae, K.T., King, B.F., Wetzel, L.H., Martin, D., Lockhart, M.E., Bennett, W.M. and Moxey-Mims, M., 2011. Potentially modifiable factors affecting the progression of autosomal dominant polycystic kidney disease. Clinical Journal of the American Society of Nephrology, 6(3), pp.640-647.

[43] Heyer, C.M., Sundsbak, J.L., Abebe, K.Z., Chapman, A.B., Torres, V.E., Grantham, J.J., Bae, K.T., Schrier, R.W., Perrone, R.D., Braun, W.E. and Steinman, T.I., 2016. Predicted mutation strength of nontruncating PKD1 mutations aids genotype-phenotype correlations

in autosomal dominant polycystic kidney disease. Journal of the American Society of Nephrology, 27(9), pp.2872-2884.

[44] Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.

[45] Breiman, Leo. "Arcing classifier (with discussion and a rejoinder by the author)." The annals of statistics 26.3 (1998): 801-849.

[46] James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.

[47] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences. 2003 Nov 24;43(6):1947-58.