USE OF SURVIVAL TREES AND RANDOM FORESTS FOR MODELING TIME TO END STAGE RENAL DISEASE

by

Alex Hurd

BS, Biochemistry, John Carroll University, 2015

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

2019

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Alex Hurd

It was defended on

April 12, 2019

and approved by

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

Jong H. Jeong, PhD, Professor and Vice Chair
Biostatistics
Graduate School of Public Health, University of Pittsburgh

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

USE OF SURVIVAL TREES AND RANDOM FORESTS FOR MODELING TIME TO END STAGE RENAL DISEASE

Alex Hurd, MS

University of Pittsburgh, 2019

Abstract

Polycystic kidney disease (PKD) is an inherited disorder distinguished by kidney cyst growth and a corresponding decline in renal function. With minimal treatment options available, often times PKD progresses to chronic kidney disease (CKD) and eventually, end-stage renal disease (ESRD), a complete loss of kidney function. The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) is a longitudinal cohort study that has been studying participants with PKD since 2001, including analysis of imaging and biomarker data. CRISP studies have characterized renal cyst growth, as well as important prognostic biomarkers for CKD. Identifying PKD patients at greatest risk for developing ESRD can help design future studies.

In this study, survival methods, including decision trees, were applied to baseline CRISP data to determine which characteristics were associated with ESRD-free survival. Results of Cox proportional hazard, survival tree, and random survival forests models showed that height-adjusted total kidney volume (htTKV), total kidney cyst volume (TCV), and glomular filtration rate (GFR) had the strongest association with ESRD-free survival.

Public health significance: Determination of which characteristics are associated with ESRD-free survival can inform preventative treatments and programs for PKD patients and give better insight to medical professionals.

Table of Contents

List of Tables

List of Figures

Preface

First and foremost, I would like to thank my thesis advisor, Dr. Landsittel, for guiding me through graduate school and for serving as a valuable resource throughout my thesis project. His constant support, encouragement, and constructive advice aided me in the completion of this thesis with relative ease. Additionally, I am grateful for Dr. Ada Youk and Dr. Jong-Hyeon Jeong for participating in my thesis committee and providing me with valuable insight.

I would also like to thank my parents, Lori and Brian, for a lifetime of support, especially for pushing me to complete extra math problems because they knew it would pay off someday. Lastly, thank you to my wife, Rachael, for bolstering me mentally, emotionally, and financially during my time as a graduate student.

1.0 Introduction

1.1 Polycystic Kidney Disease

Polycystic kidney disease (PKD), a common genetic disorder affecting roughly 500,000 people in the United States, is characterized by the development of cysts on the kidneys and a corresponding decline in renal function [1]. The most common form, autosomal dominant polycystic kidney disease (ADPKD), arises from genetic mutations in the *PKD* gene family, with roughly 80% of all cases resulting from *PKD1* gene mutations on chromosome 16, and 15% resulting from *PKD2* gene mutations on chromosome 4. The remaining few ADPKD patients have no mutations detected on *PKD1* or *PKD2* genes; therefore, a third gene, *PKD3*, has been postulated but not yet found [2]. Patients with *PKD1* mutations have faster progression of the disease, higher number of renal cysts, more severe symptoms, and lower life expectancies compared to patients with *PKD2* mutations [2, 3]. The *PKD1* and *PKD2* mutations alter the production of the proteins polycystin-1 and polycystin-2, respectively, which affect the function of the primary cilia, cellular organelles that aid in calcium detection and transport into cells [4, 5]. Ultimately, individual cells will experience increased growth in affected patients, leading to the formation of fluid-filled cysts on the kidneys [6]. ADPKD patients commonly present with hypertension, hematuria, flank pain, hepatic cysts, abdominal pain, and kidney cyst infections [4, 7]. Additionally, extrarenal complications can include hepatic cysts and brain aneurisms [8]. Symptoms commonly present between the ages of 30 to 50 [2].

Kidneys function as filters for the bloodstream by extracting waste and excreting them from the body as urine. Cysts on the kidneys inhibit this process, and the extent to which this occurs can be quantified by examining glomular filtration rate (GFR), the most common measure of renal function. Expressed in units of mL/min/1.73m^2, GFR represents the rate of blood flow through blood vessels in the kidneys, known as glomeruli, with larger values indicating greater renal function. Two primary means exist for determination of GFR; the more accurate technique is directly measuring it via iothalamate clearance, although this can be impractical and more resource-intensive. Therefore, most studies use a simpler method, estimated GFR (eGFR), that can be calculated using an equation consisting of serum creatinine level, age, sex, and race. Chronic kidney disease (CKD), an umbrella term that incorporates any kidney disease that results in long-term loss of renal function, including ADPKD, consists of five clinical stages based upon GFR values. In increasing order, the stages of CKD correspond to GFR value ranges of: > 90, 60-90, 30-60, 15-30, and < 15. End stage renal disease (ESRD), characterized by the complete loss of renal function, occurs in stage 5 CKD, as well as when patients require dialysis or kidney transplant.

Only until recently has any progress been made regarding treatment for ADPKD. A three-year study showed that the vasopressin receptor antagonist Tolvaptan is associated with slower kidney volume increases and kidney function decline [9]. Since the introduction of renal replacement therapies, namely dialysis and kidney transplant, cardiovascular diseases have become the primary cause of death in ADPKD patients [10]. Additionally, when accounting for population changes between 1996 and 2014, adjusted mortality rates for dialysis and transplant patients decreased by 32% and 44%, respectively, over that same time period [11]. However,

ESRD patients still face high mortality rates (136 per 1,000), presenting a serious health concern for the nearly 700,000 patients with ESRD in the United States [11, 12].

1.2 CRISP Study

The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) is a longitudinal cohort study that has been ongoing since 2001, with data for this analysis collected through 2017. Its purpose is to utilize techniques such as magnetic resonance (MR) imaging technology to characterize the progression of kidney growth and associated renal disease in PKD patients. A total of 241 participants diagnosed with ADPKD between the ages of 15 and 46 at baseline were enrolled in the study. Eligibility requirements included a creatinine clearance above 70 mL/min and a serum creatinine level below 1.6 mg/dL for male participants, or below 1.4 mg/dL for female participants; those with diabetes mellitus were ineligible [13]. Participants visited one of five clinical centers at baseline, followed by three annual visits, and approximately biannual visits thereafter, resulting in a total of 16 years of follow-up data at present. Data collection includes demographics, imaging, and serum and urine biomarkers, as well as estimated GFR.

Much of what is known regarding disease progression, specifically renal and cyst growth, in ADPKD patients has come from CRISP investigations. Previous studies have shown relatively stable renal function until a critical kidney volume is reached, after which a rapid decline ensues [14]. CRISP established that MR imaging techniques provide accurate and reliable measurement of renal structure in APKDP patients, found that hypertensive patients experience greater kidney cyst volume than their normotensive counterparts, and that GFR is inversely related to both kidney

and cyst volume [15]. Data from the first three years of CRISP also showed that renal and cyst growth is exponential, with an average kidney volume increase of 5.3% (204 mL) per year. Additionally, subjects with larger kidneys at baseline experienced more pronounced increases in renal volume, independent of age [16]. CRISP studies also established height-adjusted total kidney volume as an important prognostic marker for developing renal insufficiency, and showed that the type of PKD mutation affects cyst formation, but not cyst growth [17, 18].

1.3 Variables of Interest

Many of the variables being collected in the CRISP study are thought to be critical to the development of kidney disease, or at the very least relevant to its progression. Due to disparities in kidney disease prevalence and progression based upon demographics, sex, race, and age are also of interest [19-21]. CRISP findings also highlight the importance of imaging data, such as height-adjusted total kidney volume and total kidney cyst volume (htTKV and TCV, respectively) in predicting renal function over time. Additionally, given the link between PKD and hypertension, both blood pressure and hypertension status are all also relevant to disease progression. Body mass index has also been linked to various medical issues, including metabolic disorders, kidney complications, and hypertension [22-24]. Lastly, with extensive research regarding the differences in PKD1 and PKD2 prognosis, the genotype of CRISP participants was also considered. All of the variables mentioned are thought to be clinically important to the development of kidney disease and will be included in this analysis.

The CRISP study has focused primarily on using baseline data for prognosis of CDK and describing longitudinal trends in imaging biomarkers. However, the CRISP data has rarely been

framed in a time-to-event context; that is, studying time until the development of ESRD in participants, and which characteristics relate to ESRD-free survival. Until recently, time-to-event analysis has been challenging due to the inconsistent data collection time. Furthermore, participants were, by definition, healthy at baseline and often required many years of follow-up to reach ESRD. Using baseline values to predict development of ESRD years into the future is crucial due to the substantial lag time between cyst growth and renal decline. The following analysis is therefore innovative and significant.

2.0 Methods

2.1 Survival Analysis

In survival analysis participants are followed until they reach some event of interest, such as death or disease progression, or are censored, i.e. removed from the study because of events unrelated to the study or being lost to follow-up. Survival data is visualized by plotting a survival function, $S(t)$, over time to quantify the probability of an individual surviving, or not experiencing the event, beyond time t. Mathematically, this can be shown as follows:

$$S(t) = P(T > t) \tag{1}$$

Survival curves are typically displayed using the non-parametric Kaplan-Meier product-limit estimators. At each time point, the probability of survival is calculated using the product of conditional probabilities:

$$\hat{S}(t) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right) \tag{2}$$

where d_i is the number of events that have occurred by time t_i, and n_i is the number of subjects at risk at time t_i. Censoring is accounted for between events and the calculation is straightforward, making it ubiquitous in displaying survival functions [25]. By definition, at time $t = 0$, no event has occurred, and $\hat{S}(t) = 1$, and the Kaplan-Meier estimate is a non-increasing step function.

To evaluate between-group differences in survival times, the log-rank test compares the observed and expected number of events at each event time. The test statistic is calculated using equation 3,

$$Q = \frac{\left[\sum_{i=1}^{I}(d_{i1} - e_{i1})\right]^2}{\sum_{i=1}^{I} V_i} \tag{3}$$

where d_{j1} is the number of events that have occurred by time t_i in group 1, which follows a hypergeometric distribution with mean and variance:

$$e_{i1} = d_i\left(\frac{n_{i1}}{n_i}\right) \tag{4}$$

$$V_i = d_i\left(\frac{n_i - d_i}{n_i - 1}\right)\left(\frac{n_{i1}}{n_i}\right)\left(1 - \frac{n_{i1}}{n_i}\right) \tag{5}$$

Under the null of equal survival, the test statistics, Q, follows a chi-square distribution with one degree of freedom. Log-rank tests can be extended to multiple groups, where the statistic for k groups follows a chi-square distribution with k-1 degrees of freedom. Additionally, the test of trend can be used to evaluate ordered alternatives, against a chi-square distribution with one degree of freedom [25].

Another common tool utilized in survival analysis is the Cox proportional hazards model. The hazard function, $h(t)$, represents the probability of experiencing the event, given it has not occurred up until time t. The model assumes that the risk, or hazard, of experiencing the event is independent of time [25]. The proportional hazards assumptions states that the ratio of the hazard

function to the baseline hazard is constant over time [26]. The model incorporates multiple variables using the form:

$$h(t;x) = h_0(t)e^{\beta x} \tag{6}$$

where $h_0(t)$ is the baseline hazard, x is a covariate, and β is the effect of a one unit change of x on the log-hazard. The baseline hazard represents the hazard function when all the covariate equals zero. Extending the model to p covariates gives the form:

$$h\left(t;x_1,\dots,x_p\right) = h_0(t)e^{\beta_1 x_1+\dots+\beta_p x_p} \tag{7}$$

A positive value of β equates to e^{β} above one, i.e. an increase in hazard; conversely, negative β values correspond to a decrease in hazard. Integrating the hazard function yields the cumulative hazard function, $H(t)$, a measure of the accumulated risk someone faces over a period of time:

$$H(t) = \int_0^t h(u)\,du \tag{8}$$

The resulting function starts at zero and increases over time as a non-decreasing step function. The cumulative hazard and survival functions are related through the following equations:

$$H(t) = -\ln S(t) \tag{9}$$

$$S(t) = e^{-H(t)} \tag{10}$$

This implies that, given a constant hazard of an event occurring, the survival function will decrease exponentially over time. Constant hazard can be explicitly tested by analyzing the relationship

between the residuals and time, either by plotting the two or conducting a global Schoenfield test [26]. If the model contains k variables, this test follows a chi-square distribution with k-1 degrees of freedom; significant p-values indicate a violation of the constant hazard assumption.

2.2 Decision Trees

Classification and regression trees (CART), as proposed by Breiman, models the relationship between an outcome and a list of predictors by recursively partitioning the data. When the outcome is continuous, regression trees are used; similarly, binary outcomes correspond to classification trees. Beginning with a constant that is equivalent to the mean outcome value, the data is split into two groups, each of which is given its own constant value (Figure 1). Splits are performed repeatedly until certain stopping criteria are met, typically when a group size falls below a pre-determined value [27]. Eventually, the entire function will consist of a series of segments approximating the true model (Figure 2).

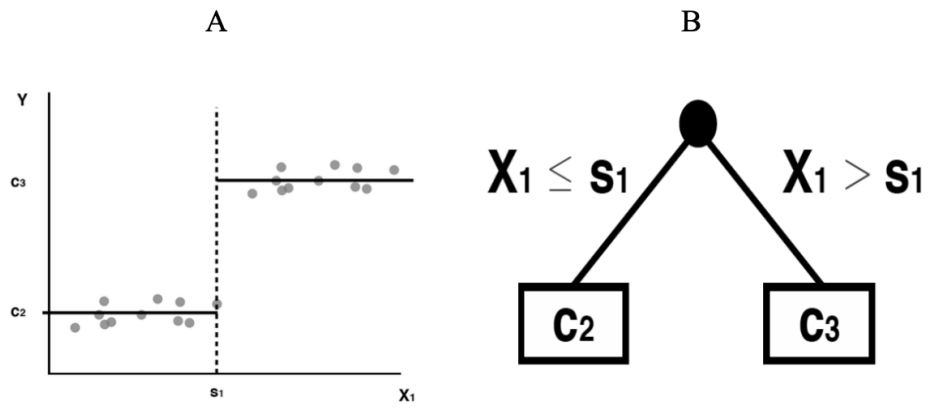

Figure 1. Visualization of decision trees

(A) Splitting the data at point s_1 (B) Resulting tree split into two constants, c_2 and c_3

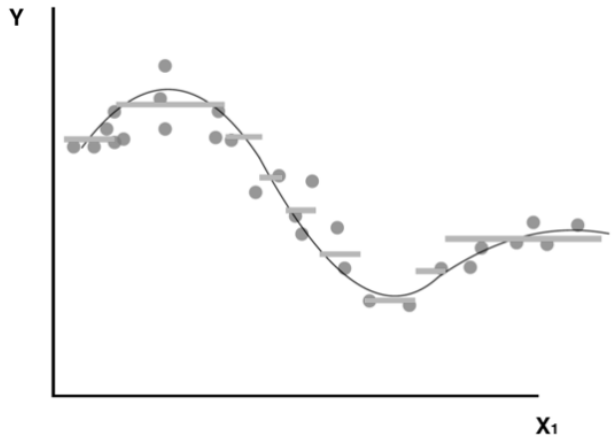

Figure 2. Example of decision tree final model

Trees can be extended to multiple covariates, but it is much more difficult to conceptualize in a visual manner.

Decision trees attempt to create increasingly homogenous groups in terms of the outcome based on some measure of node impurity [28]. Node impurity gives a measure of how much variation in the outcome exists within a node, with higher values indicating more impure, and therefore less homogenous, nodes. For regression trees and continuous outcomes, the node impurity measure used is mean squared error (MSE), which quantifies the average squared difference between each data point and the predicted outcome for that subgroup, or average outcome value, and are summed across all subgroups. For a given subgroup, the MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \quad (11)$$

where y_i is the value of the outcome for the i^{th} observation in the subgroup and $\hat{y}$ is the average value of the outcome of that subgroup. If there are n data points, there exist $n-1$ possible split points, $s_i = \{s_1, s_2, \ldots, s_{n-1}\}$, each with a corresponding MSE. The MSE at each split consists of the both the left and right components:

$$MSE_{total} = MSE_{left} + MSE_{right} \tag{12}$$

The tree algorithm identifies the split which minimizes the total MSE, creating two regions. This process is then repeated, finding the best possible split point in every region, until some stopping rule is satisfied, typically once a terminal node reaches a small enough size.

For binary or multinomial outcomes, the classification tree uses the Gini index to measure node impurity:

$$G = \sum_{k=1}^{K} p_k(1 - p_k) \tag{13}$$

where p_k is the proportion of observations from the k^{th} class. The Gini index can be written as:

$$G = 1 - P(k = 0)^2 - P(k = 1)^2 \tag{14}$$

Node purity is maximized when the Gini index is minimized (i.e. a Gini index of 0 results from perfect classification).

Decision trees have the potential to create complex, high-variance models that overfit the data by creating too many partitions. In order to lower the variance at the cost of marginally increasing the bias, the overall size of the tree can be decreased through pruning [27]. Once a tree

is fully grown, the least helpful splits are pruned back by reversing that split's partition of data. Consideration of where to prune on a tree is determined by the largest reduction in the overall model's performance criteria, MSE or Gini index.

2.3 Survival Trees

Tree-based models can be extended to survival data, where the covariate space is recursively partitioned based on outcome values, albeit with different splitting criterion. In order to minimize variations within groups and maximize variations between groups, observations are clustered into nodes by median survival time [29, 30]. Splitting criterion is based on finding the most significant log-rank statistic, ensuring the best separation of median survival times between the two resulting nodes [31]. Censoring must be taken into account when determining the size of a survival tree. Growth will stop when a node reaches a specified size, all members of a node are censored, or all members have identical covariates [32].

Survival trees involves pruning to prevent over-fitting by incorporating established survival analysis methods. As described by Paravati et al, an algorithm for consolidating nodes based on pairwise log-rank tests can reduce the number of terminal nodes in a survival tree:

1) Order all groups by median survival time
2) Carry out log-rank tests between adjacent groups
3) If two groups have non-significant test results, combine those nodes into one group.
4) Recalculate survival curves and median survival times
5) Repeat until no more groups can combine

The advantage of this method is two-fold: firstly, it is a variant of pruning that prevents overfitting of the data. Secondly, it can effectively "rank" the groups by survival time, a useful and easily interpretable result that allows clinical investigators to delineate between various risk classes. For instance, the group with the lowest median survival time can be designated as high risk, while the group with the highest median survival time can be designated as low risk [32].

Compared to traditional survival analysis methods, survival trees allow for a more robust model, and have the advantage of relaxing some of the underlying restrictive assumptions, notably the constant hazard assumption [33]. They allow for easy interpretability, particularly in a clinical context, by sorting patients' prognosis based on covariate values, as well as being able to ascertain complex interactions between covariates.

2.4 Ensemble Methods

Ensemble methods combine results from individual models, known as base learners, to increase predictive power. Random forest, also proposed by Breiman, use decision trees as base learners, with the intent to build upon their strengths while simultaneously enhancing their weaknesses. A single tree can be greatly influenced by statistical noise and have high variance, even after pruning. Due to the adaptive nature of decision trees, small changes in the data used to grow the tree can drastically alter the tree model, making it relatively unstable [34].

Random forests incorporate multiple models to lower the overall variance of the model by averaging their results. Randomness is integrated into the process through a two-fold procedure. Firstly, a bootstrap sample for each tree in the forest is drawn from the data; that is, if a dataset has

n observations, a sample of size n is drawn, with replacement. On average, roughly two-thirds of the original dataset is included in the bootstrap sample, with the other third, referred to as the out-of-bag (OoB) sample, being excluded. This process is executed repeatedly until the desired number of bootstrap samples has been established. Secondly, at each node, a random subset of predictors is chosen, and the best split comes from only those variables. If there exists p predictors in the model, then $d = \sqrt{p}$ (classification) or $d = p/3$ (regression) predictors are randomly selected at each node. For every single split, a new random set of d predictors will be selected for consideration [34]. The random forest algorithm is described below:

1) Randomly draw a bootstrap sample of size n from a dataset with n observations. Repeat B number of times. Each bootstrap sample will be used for a different tree.
2) At each node, randomly select $d < p$ subset of predictors for splitting consideration.
3) Repeat until the tree is fully grown; that is, until a stopping rule prevents any further partitions.
4) Do not prune the trees, leave them fully grown in order to utilize the maximum number of splits.

The strength of random forests lies in both the high number of trees used, and the randomness introduced into the model, which ensures the trees are decorrelated by growing independently of a constant set of predictors. When a single, strong predictor dominates a tree model, it often is the first split in that tree. Removing a majority of the predictors at the first, and every subsequent, split allows the trees to perform differently [28].

Random forest accuracy is evaluated using OoB samples, with each tree in the forest being tested by its respective OoB sample. For all possible pairs in the OoB sample, the observed and predicted outcomes are compared. If the outcomes are concordant, that pair receives a value of one, discordant pairs receive a value of zero, and ties receive a value of one-half. Summing these values and dividing by the total number of pairs gives the C-index. OoB prediction error is defined as 1 – C-index, with lower values indicating higher predictive power [34]. Variable importance measures summarize how critical each variable is in terms of partitioning the data and its predictive power, and is calculated by removing that variable from the analysis and observing the mean decrease in accuracy compared to the model with the variable included [34]. Relative variable importance scales these values between zero and one, with the highest variable importance measure serving as the reference with a value of one.

Ensemble methods applied to survival trees gives random survival forests, which operate similar to traditional random forests. When calculating OoB prediction error, the observed outcome is survival time, and variable importance measures are calculated in a similar fashion. To interpret, the higher the relative variable importance, the more that variable contributes to maximizing the log-rank statistic between nodes [33].

3.0 Results

3.1 Data Overview

Out of the 241 CRISP participants, four were excluded due to lack of necessary data. Of the remaining 237 participants, 50 experienced ESRD (21%); three died during the study and were censored. All 184 participants who did not experience ESRD or death during the study were censored at the end of follow-up.

The following variables were considered during the analysis: age, gender, race, body mass index (BMI), glomular filtration rate (GFR), height-adjusted total kidney volume (htTKV), total kidney cyst volume (TCV), genotype (PKD1 mutation vs. PKD2 mutation / no mutation detected), hypertension status (Y/N), and the average of seated and standing blood pressure (both diastolic and systolic). Baseline measurements were used for all variables.

Summary statistics for continuous variables are given for all participants (Table 1), including an overall summary, as well as divided by ESRD status. Boxplots display the same information for a visual comparison (Figure 3). Frequency and percentage of each level for categorical variables was calculated (Table 2). The number and percentage of ESRD cases at each level is also given.

Table 1. Descriptive statistics for continuous variables

Variable	Mean	SD	Min	Median	Max
Age (years)	32.40	8.81	14.78	33.85	46.30
No ESRD	31.43	9.09	14.78	32.30	45.97
ESRD	26.02	6.57	22.48	26.07	46.31
Diastolic (mmHg)	79.26	10.46	53.56	79.33	109.56
No ESRD	78.15	10.27	53.56	78.33	107.33
ESRD	83.41	10.24	62.22	80.83	109.56
Systolic (mmHg)	123.05	12.91	91.00	123.56	157.33
No ESRD	122.24	13.00	91.00	122.83	157.33
ESRD	126.07	12.19	101.44	125.06	154.22
BMI (kg/m^2)	25.68	5.27	17.26	25.15	50.90
No ESRD	25.55	5.35	17.26	24.65	50.90
ESRD	27.03	4.85	18.48	26.56	41.35
GFR ($mL/min/1.73m^2$)	92.60	22.92	49.69	90.62	173.46
No ESRD	97.23	22.00	49.69	93.82	173.46
ESRD	75.28	17.50	50.14	76.78	130.00
htTKV (cc/m)	621.38	374.15	167.98	504.41	2113.12
No ESRD	512.41	274.09	167.98	432.31	2113.12
ESRD	1028.27	417.48	341.43	976.95	1915.17
TCV (cc)	534.63	529.69	2.39	338.47	2971.12
No ESRD	386.341	392.58	2.39	262.72	2971.12
ESRD	1089.27	607.01	149.44	967.76	2353.86
Follow-up (months)	130.78	39.71	13.45	149.91	183.33
No ESRD	132.78	41.31	13.45	152.02	180.27
ESRD	123.30	32.35	35.64	126.53	183.33

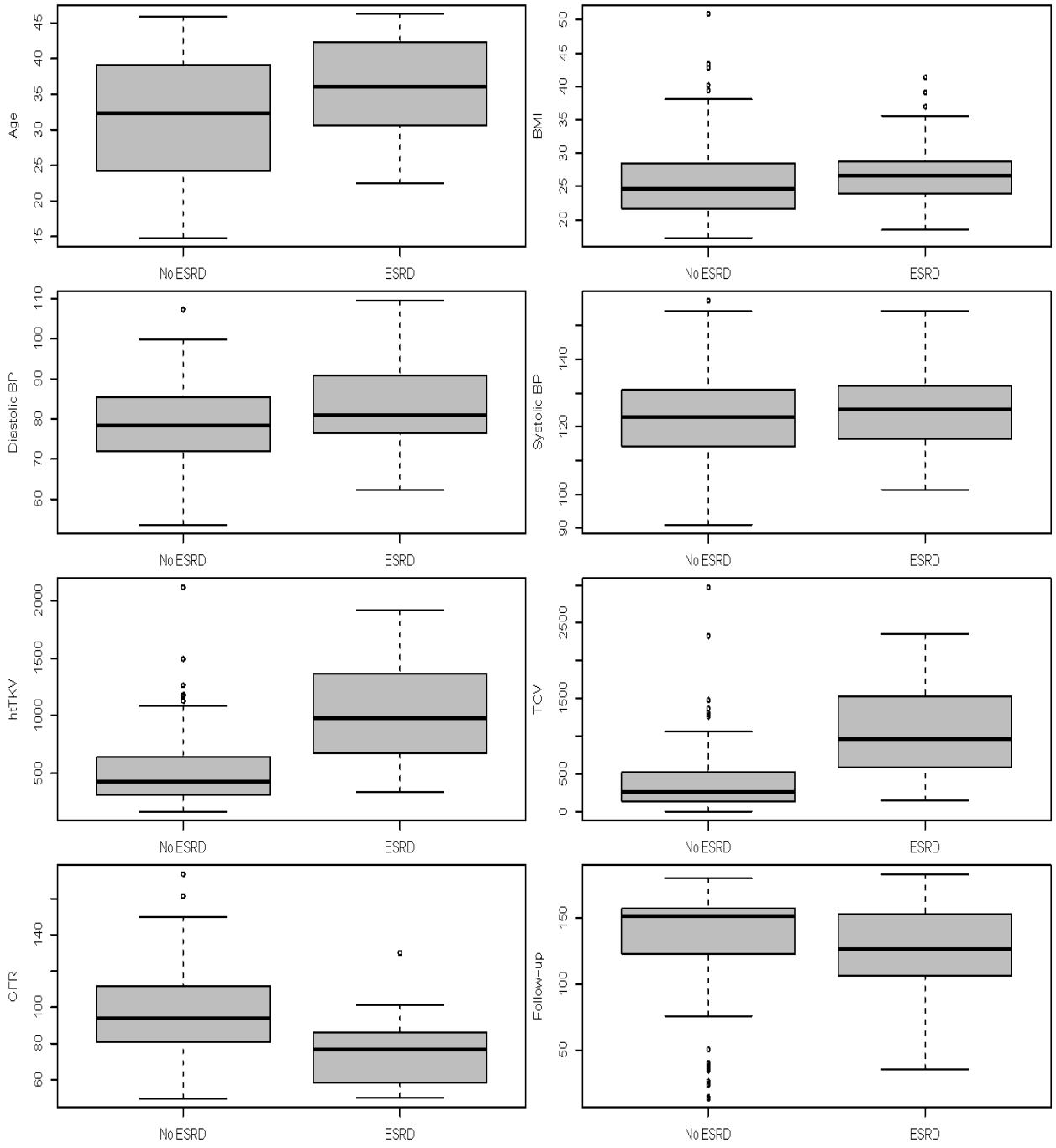

Figure 3. Distribution of continuous variables, by ESRD status

Table 2. Descriptive statistics for categorical variables

Variable	Frequency	Percentage	ESRD #	ESRD %
Gender				
Male	95	40.08	22	23.16
Female	142	59.92	28	19.72
Genotype				
PKD1	185	78.06	47	25.41
PKD2 / NMD	52	21.94	3	5.77
Hypertension				
Yes	144	60.76	44	30.56
No	93	39.24	6	6.45
Race				
Caucasian	208	87.76	45	21.63
Afr. American	25	10.55	4	16.00
Asian	2	0.84	0	0.00
Hispanic	2	0.84	1	50.00

Table 1 shows that most continuous variables in the study are normally distributed, with the exception of htTKV and TCV, both of which are right-skewed. While most survival outcome data are right-skewed, the left-skew of follow-up months seen in Table 1 signals that more events happen later in the CRISP study, which is consistent with what is known about the slow progression of PKD. Comparing boxplots in Figure 3 reveals that participants who experience ESRD have higher values of age, BMI, diastolic and systolic blood pressure, htTKV, and TCV, as well as lower GFR and follow-up duration, compared to those who have not experienced ESRD. The largest differences based on ESRD status exist for htTKV and TCV. Table 2 highlights the differences in ESRD occurrence between the levels of categorical variables. Higher ESRD prevalence is seen in men, Caucasians and Hispanics, PKD1 genotypes, and hypertensive participants, the latter two results being consistent with what is known regarding PKD. Hypertension status has the largest discrepancy in ESRD prevalence.

A Kaplan-Meier survival curve (Figure 4) fitted to the CRISP data confirms the increasing prevalence of ESRD as the study continued, implied by the left-skew of the follow-up variable.

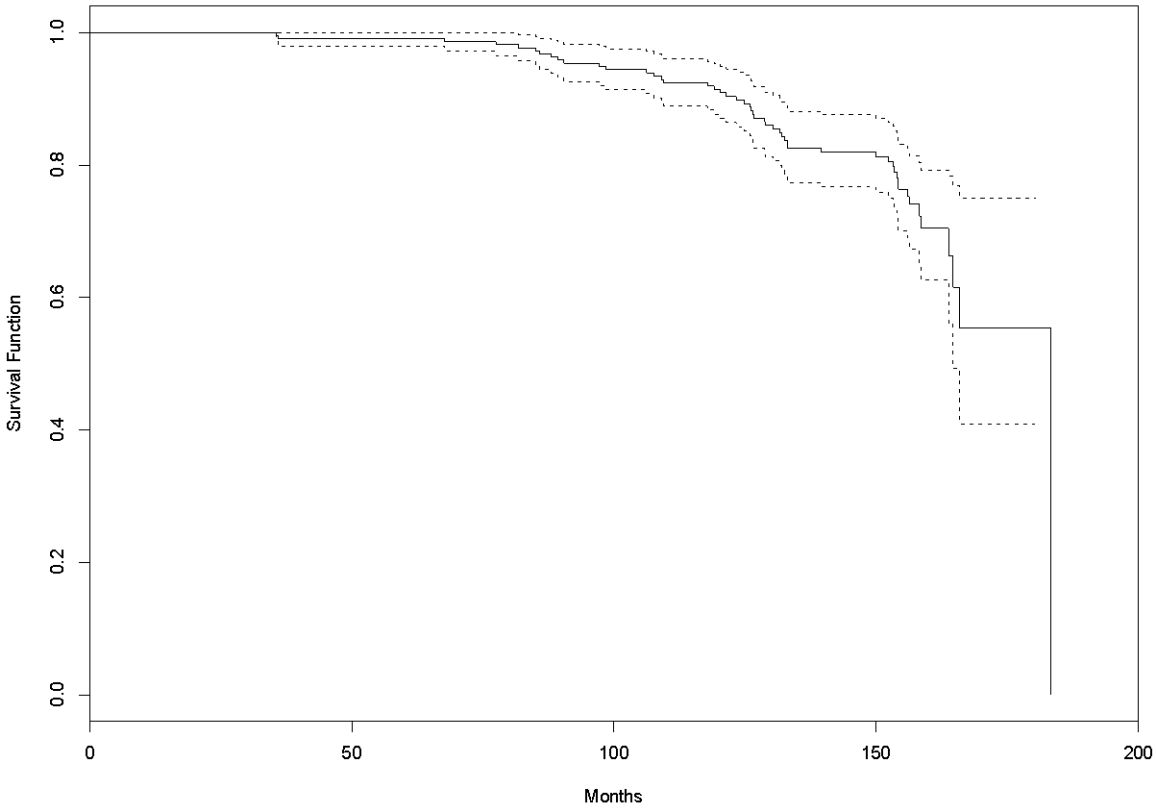

Figure 4. Overall Kaplan-Meier survival curve for ESRD, with 95% confidence interval

3.2 Cox Proportional Hazard

Unadjusted Cox proportional hazard models were fit using each variable individually to determine the marginal effect of all variables on the hazard of ESRD. A model containing ten of the 11 variables was fit to determine the adjusted effect on hazard. TCV was excluded due to its high correlation with htTKV. Results from both models were combined (Table 3). An additional model with TCV that excludes htTKV is shown in Appendix A (Table 7).

Table 3. Summary of Cox proportional hazards model

Variable	Unadjusted			Adjusted		
	β	e^{β}	P-value	β	e^{β}	P-value
Age	0.068	1.070	< 0.001	0.003	1.003	0.911
Gender	-0.160	0.852	0.58	-0.158	0.854	0.626
Race	-0.099	0.905	0.75	0.416	1.516	0.208
BMI	0.043	1.044	0.061	0.014	1.014	0.641
Diastolic	0.039	1.040	0.005	0.027	1.027	0.301
Systolic	0.017	1.017	0.12	-0.025	0.975	0.224
Hypertension	1.618	5.043	< 0.001	0.678	1.969	0.150
GFR	-0.054	0.948	< 0.001	-0.036	0.964	< 0.001
htTKV	0.002	1.002	< 0.001	0.001	1.001	< 0.001
TCV	0.001	1.001	< 0.001	---	---	---
Genotype	-1.457	0.233	0.015	-1.137	0.321	0.078

Using the unadjusted models, seven variables (age, diastolic, hypertension, genotype, GFR, htTKV, and TCV) were significant ($p < 0.05$), with the latter three being extremely significant ($p << 0.05$). When combining into one model, age, diastolic, and genotype were no longer significant. Based on these results, GFR, htTKV, and TCV are all very significantly associated with the hazard of developing ESRD.

Positive coefficients are associated with an increase in hazard, meaning ESRD risk is related to increases in age, BMI, diastolic blood pressure, systolic blood pressure, TCV, and htTKV, as well as having hypertension, being white or being male, and having the PKD1 genetic mutation; conversely, ESRD risk relates to decreases in GFR. However, when moving from the unadjusted to adjusted model, systolic blood pressure converted from having a positive association with ESRD hazard to a negative association.

To check the constant hazard assumption, a cumulative hazard function was plotted over time (Figure 5). The exponential shape indicates that hazard increases over time and may not be

constant. To formally test this, a global Schoenfield test was conducted (χ^2 = 14.2, p = 0.1628), indicating that no violation of the constant hazard assumption occurred. While the curve does not appear to show constant hazard, the formal test confirms that this is, in fact, true.

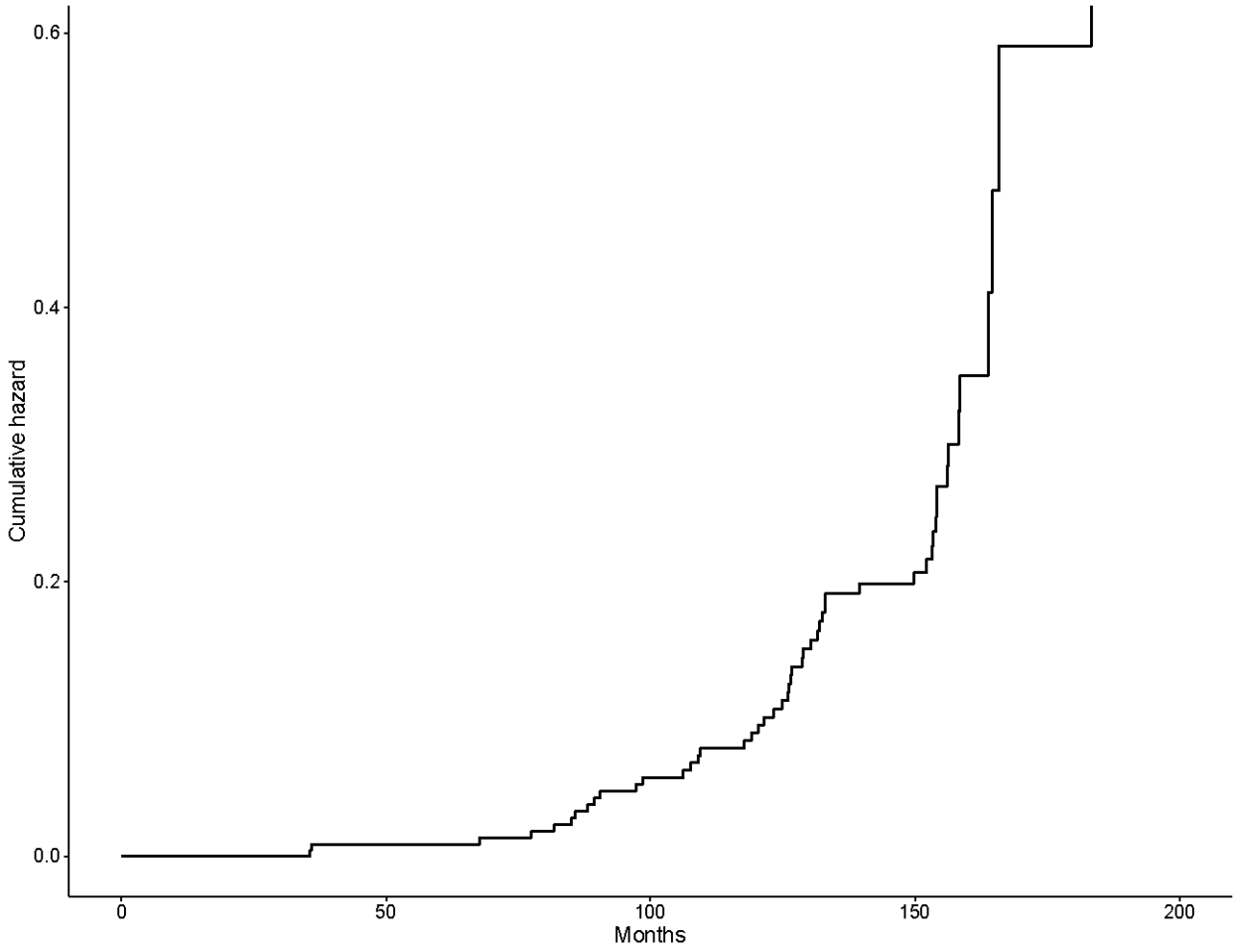

Figure 5. Overall cumulative hazard curve

3.3 Survival Tree

An unpruned survival tree was fit to the CRISP data (Figure 6), with splits occurring on htTKV, TCV (twice), GFR, BMI, and diastolic blood pressure (twice); age, gender, genotype, hypertension status, and systolic blood pressure did not appear in the model. Each of the eight terminal nodes had a corresponding survival function for the subjects in that specific node, shown at the bottom of Figure 6, with the y-axis as survival rate and the x-axis as follow-up time, in months.

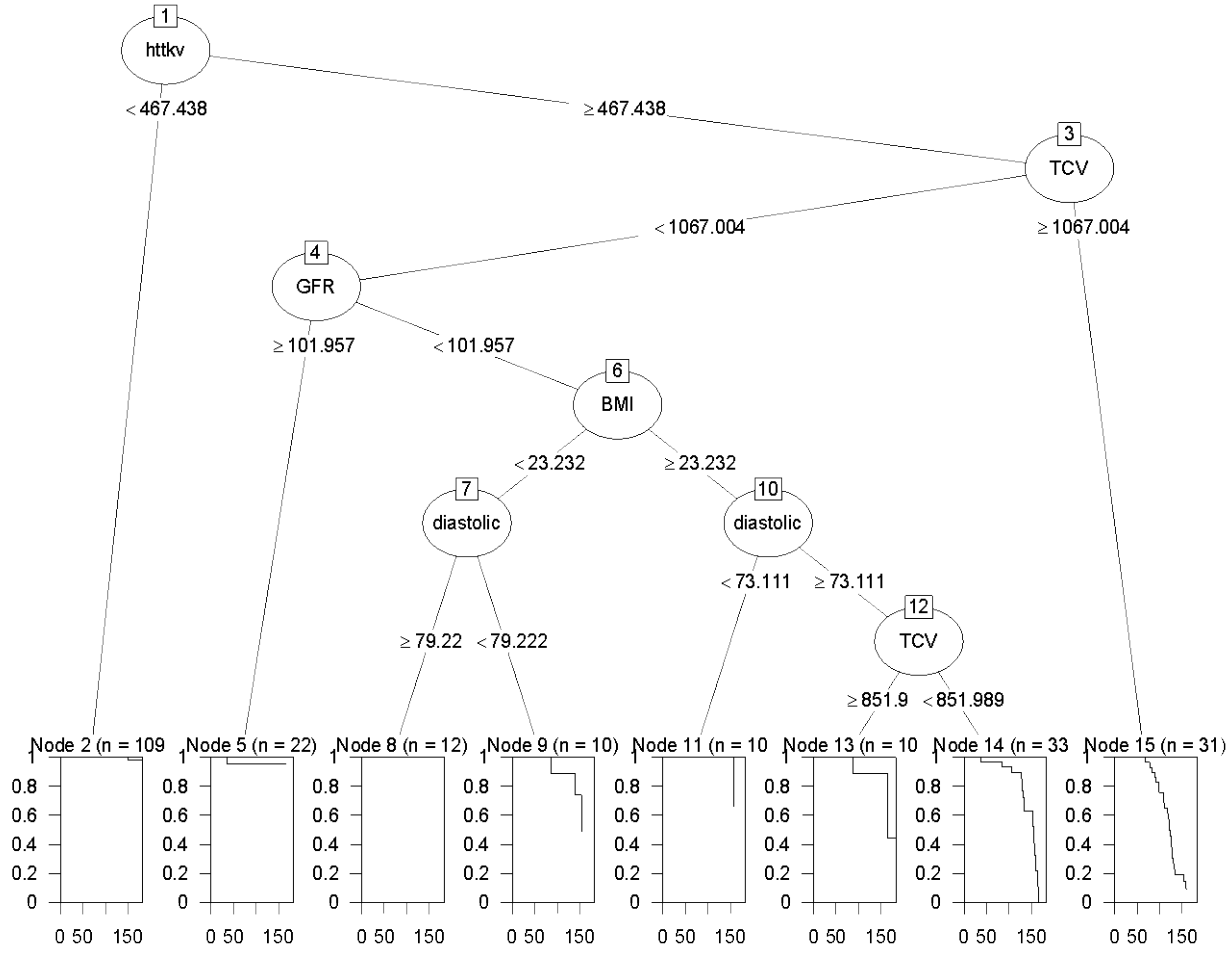

Figure 6. Survival tree model without pruning

Summary information of the terminal nodes are displayed in Table 4. The survival curves from the eight terminal nodes were combined into one plot to for a visual comparison (Figure 7).

Table 4. Summary of survival tree by node

Node	Median Survival	Size	ESRD #	ESRD %
2	153.50	109	1	0.92%
5	151.78	22	1	4.55%
8	154.78	12	0	0.00%
9	144.90	10	3	30.00%
11	156.07	10	1	10.00%
13	151.79	10	3	30.00%
14	131.58	33	17	51.52%
15	119.06	31	24	77.42%

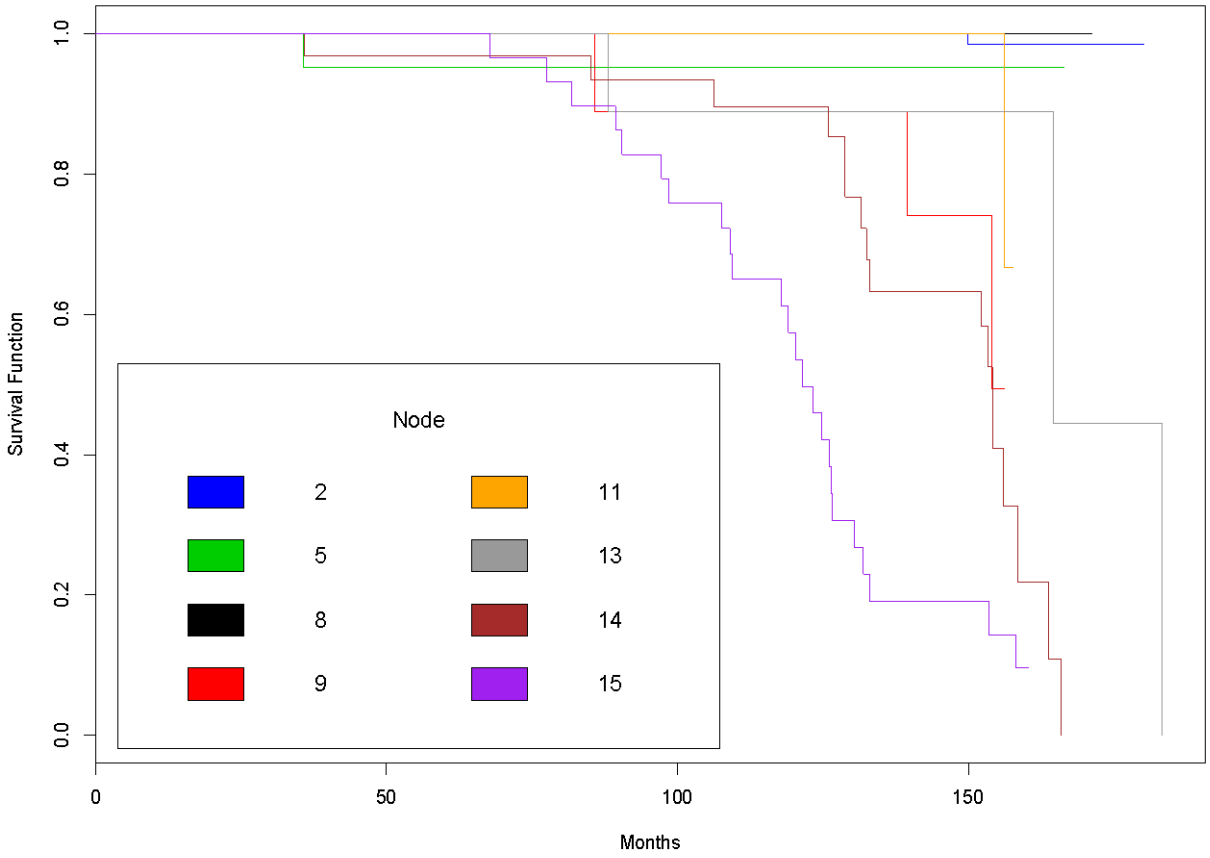

Figure 7. Kaplan-Meier survival curves by node

Based on the median survival time and survival functions, marked differences exist between the nodes. However, certain nodes have extremely similar survivals, and may not need to be in separate groups. In order to prevent overfitting of the data, the log-rank algorithm for node consolidation was performed. Following five rounds of consolidating groups, only three groups remained, as shown in the final, pruned survival tree model (Figure 8). These three groups were put into classes based on prognosis, with class I corresponding to the best overall prognosis, and class III corresponding to the worst overall prognosis. Summaries of the prognostic classes (Table 5) and survival functions (Figure 9) show distinct survival functions between classes. Nodes 2, 5, 8, and 11 were consolidated into class I, nodes 9, 13, and 14 were consolidated into class II, while node 15 became class III.

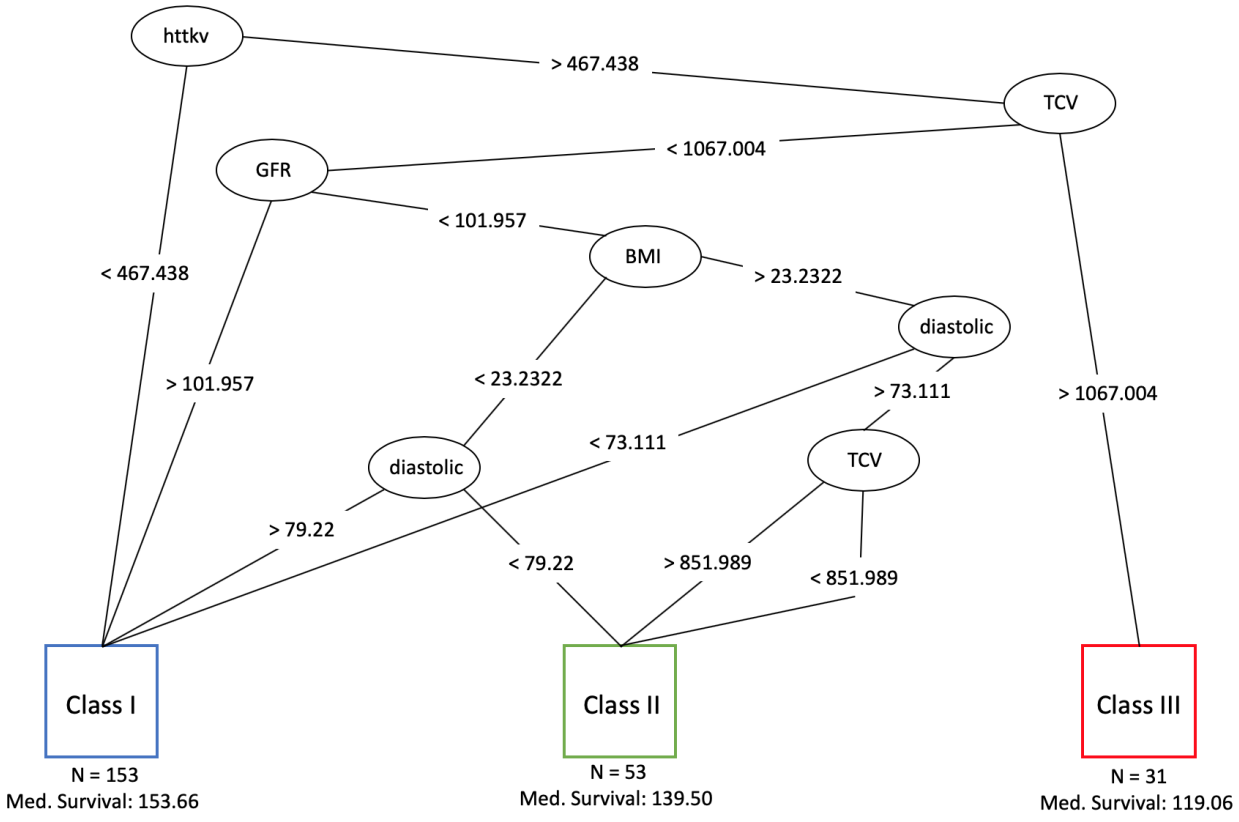

Figure 8. Survival tree model with pruning

Table 5. Summary of pruned tree by prognostic class

Class	Median Survival	Size	ESRD #	ESRD %
I	153.66	153	3	1.96
II	139.50	53	23	43.40
III	119.06	31	24	77.42

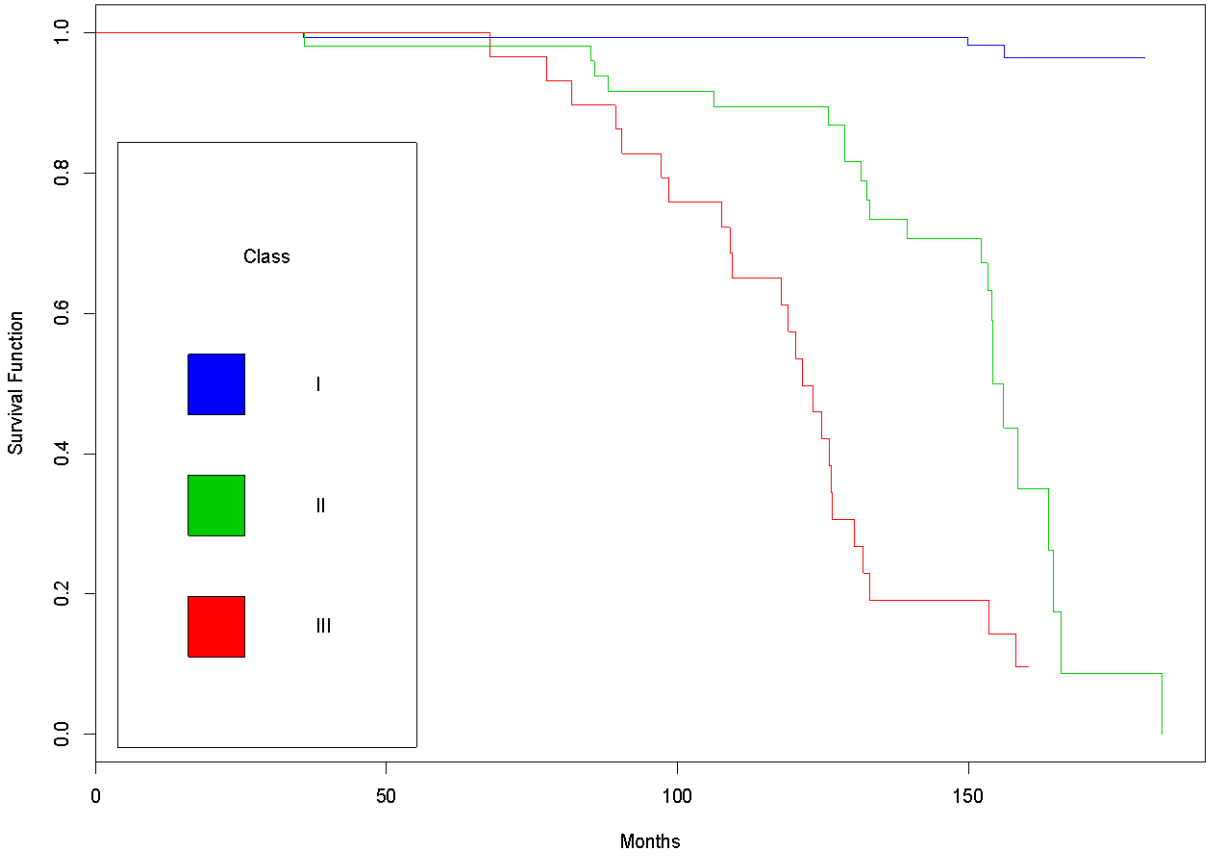

Figure 9. Kaplan-Meier survival curves by prognostic risk class

The pruned survival tree partitions CRISP participants into three groups with distinct survivals. Those in class I seldom experience ESRD, whereas a majority of class III subjects experience ESRD. This verifies that the covariates used in the model can be used to differentiate

the prognosis of participants using baseline measurements. Notably, imaging data are involved in the first two splits of the survival tree, implying that they contribute most to placement into prognostic classes.

3.4 Random Survival Forest

A random survival forest was created using 500 bootstrap samples of size 237. Four variables were randomly considered at each split; the average number of terminal nodes for the forest was 12.55. Variable importance and relative importance measures were calculated and are tabulated (Table 6) and displayed (Figure 10) below. The out of bag (OoB) error rate was measured at every forest tree size (Figure 11).

Table 6. Variable importance for random survival forest

Variable	Importance	Relative Importance
htTKV	0.1009	1.0000
TCV	0.0799	0.7915
GFR	0.0315	0.3120
Hypertension	0.0067	0.0666
Age	0.0033	0.0330
Diastolic	0.0023	0.0226
Systolic	0.0020	0.0195
Genotype	0.0018	0.0178
BMI	0.0002	0.0022
Gender	0.0001	0.0009
Race	-0.0002	-0.0022

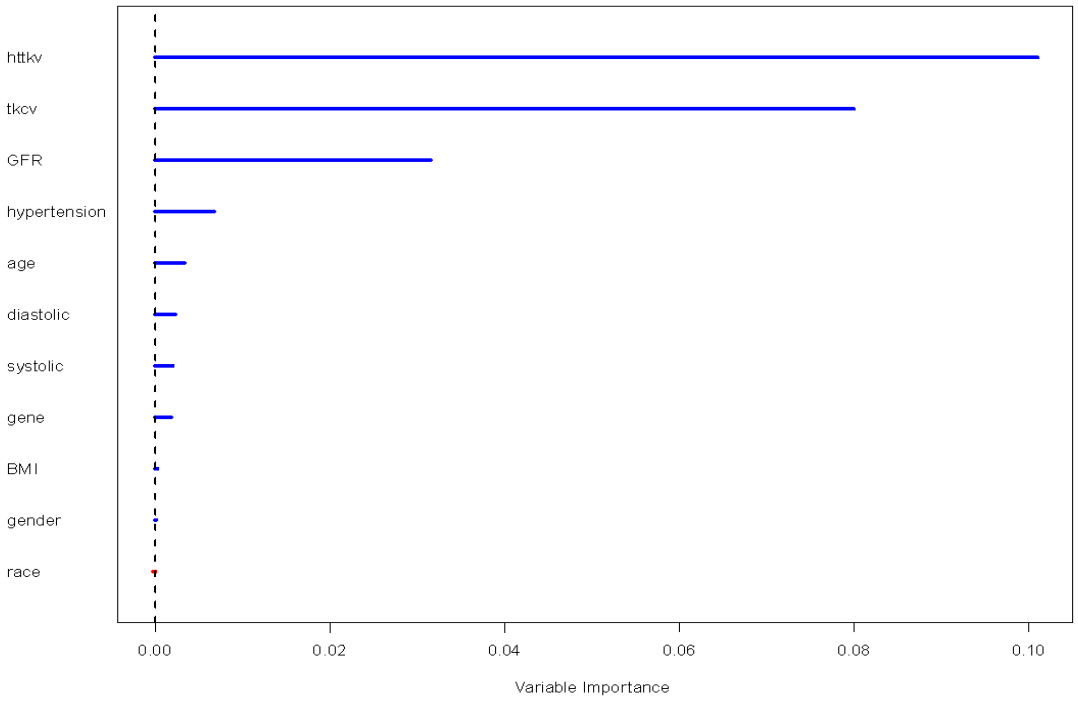

Figure 10. Visual display of variable importance for random survival forest

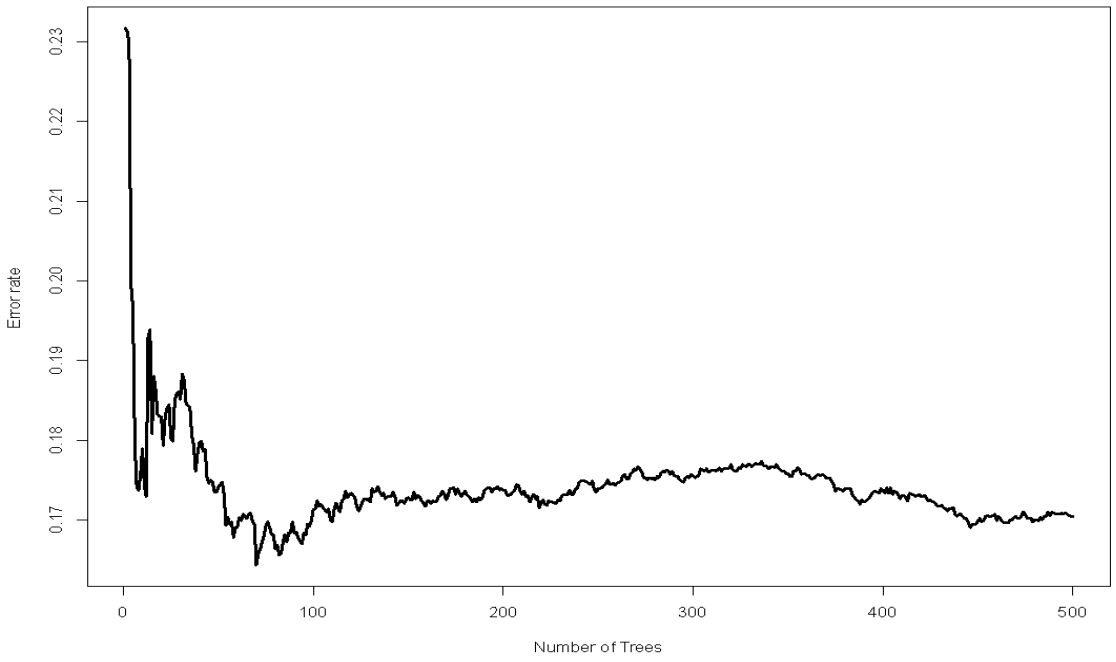

Figure 11. Out of bag error rate as a function of random survival forest size

The OoB error rate stabilizes to 16.86% by 500 trees, but saw minimal decreases after 75 trees, suggesting the addition of trees beyond this number is computationally unnecessary. A single tree in a random forest sees the error rate increase by 6.78% to 23.64%. Variable importance measures indicate that htTKV, TCV, and, to a lesser extent GFR, contain the most predictive ability of all variables in the analysis by a wide margin. The marginal effect of each variable on predicted survival was calculated and plotted. By controlling for all other variables, they show each variable's relation with predicted survival. These can be seen in Appendix A (Figures 12-22).

4.0 Discussion

Summaries of the variables indicate imaging data most closely relate to ESRD status. A Cox proportional hazards model confirms this through the highly significant associations htTKV, TCV, and GFR have with ESRD hazard. Due to the high collinearity of htTKV and TCV, having only one of these variables in the model at a time limits how the Cox model can be interpreted. This information alone does not resolve exactly how, or to what extent, these variables are associated with ESRD risk.

Using survival tree models not only allows the addition of correlated variables, but also further clarifies the associations seen between the variables and ESRD status. TCV values greater than 1067 cc at baseline result in placement into the high-risk class III, with over three-fourths of all participants experiencing ESRD within 16 years of follow-up, while htTKV values less than 467 cc places a participant into the low-risk class I. Simple baseline imaging measurements can greatly inform clinicians about the prognosis of PKD patients. Previous CRISP studies established imaging measurements, particularly htTKV, as significant biomarkers for developing ESRD. This study further explores the nature of this relationship by highlighting specific values that are important in determining prognosis. Additionally, the complex interactions at play between the variables allows for a robust method of sorting participants, coupled with easy interpretability.

When expanded to a random survival forest with 500 trees, htTKV, TCV, and GFR remained the most crucial variables based on relative importance measures, with all other variables only minimally contributing to the model's predictive ability. The 6.78% decrease in error rate from a single tree to a forest model, while small, does illustrate that utilizing random survival forests increases the predictive power compared to a single survival tree.

All three models contributed useful information concerning between development of ESRD and the variables in the study. The Cox proportional hazards model established which variables are associated with increased or decreased hazard. The survival tree model revealed how all covariate values can produce a general model that predicts prognosis of PKD. The random survival forest model further explains which variables are most crucial in predicting ESRD.

In summary, the survival models in this study focus on htTKV, TCV, and GFR baseline values as being most pertinent in explaining PKD prognosis because of their strong association with ESRD-free survival. This information can be used for early detection and prognosis of PKD, before renal function declines, allowing for more time to focus on treatment and prevention. With so few cases of ESRD in the CRISP study, additional years of follow-up will further elucidate the relationship between the covariates and development of ESRD in PKD patients.

Appendix A: Supplemental Material

Table 7. Adjusted Cox model results with TCV included and htTKV excluded

Variable	Unadjusted			Adjusted		
	β	e^{β}	P-value	β	e^{β}	P-value
Age	0.0680	1.0704	0.00024	0.0022	1.0022	0.93034
Gender	-0.160	0.852	0.58	-0.1694	0.8442	0.60465
Race	-0.0994	0.9054	0.75	0.4535	1.5738	0.16723
BMI	0.0428	1.0437	0.061	0.0193	1.0195	0.50739
Diastolic	0.0393	1.0401	0.0047	0.0264	1.0267	0.30989
Systolic	0.0171	1.0173	0.12	-0.0238	0.9765	0.24965
Hypertension	1.618	5.043	0.00021	0.7348	2.0851	0.11994
GFR	-0.0539	0.94752	6.5e-10	-0.0401	0.9607	0.00011
htTKV	0.0021	1.0022	2e-16	---	---	---
TCV	0.0012	1.0012	3.2e-16	0.0007	1.0007	0.00102
Genotype	-1.457	0.233	0.015	-1.2663	0.0.2819	0.04913

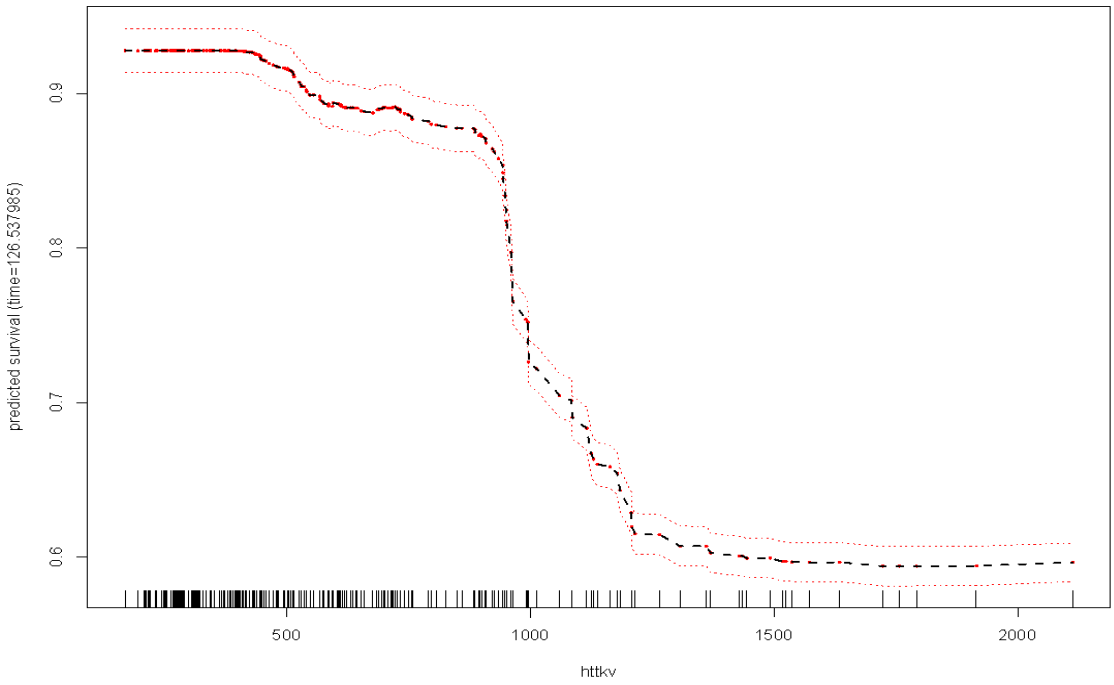

Figure 12. Marginal effect of htTKV on predicted survival

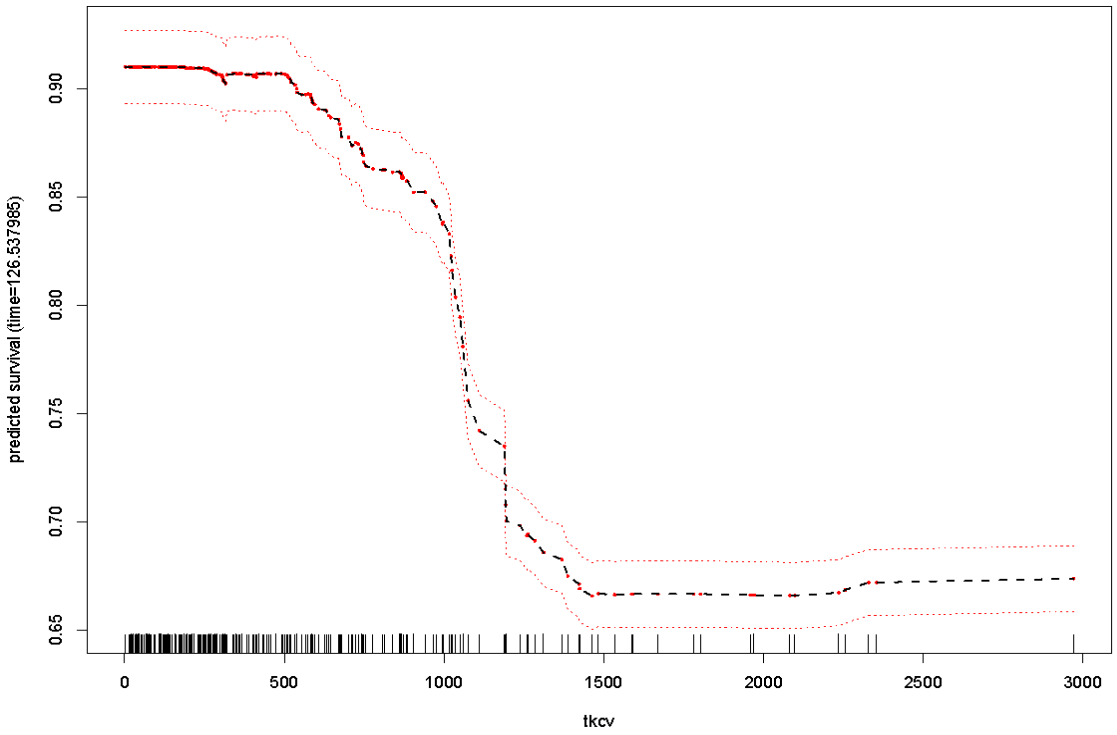

Figure 13. Marginal effect of TCV on predicted survival

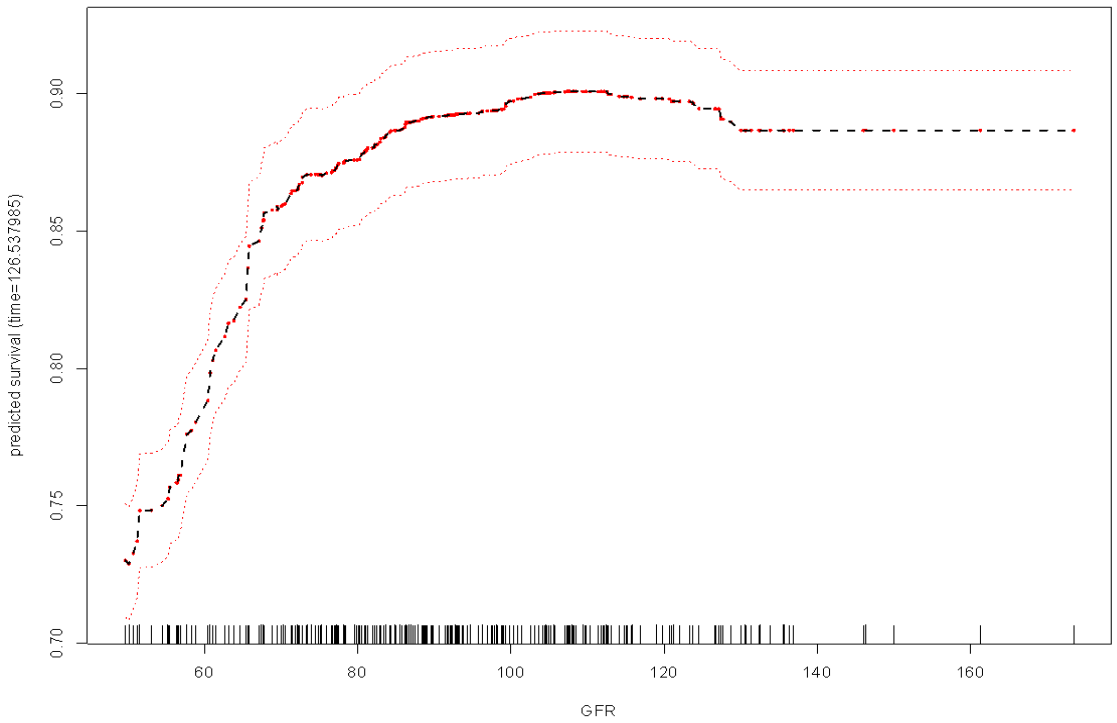

Figure 14. Marginal effect of htTKV on predicted survival

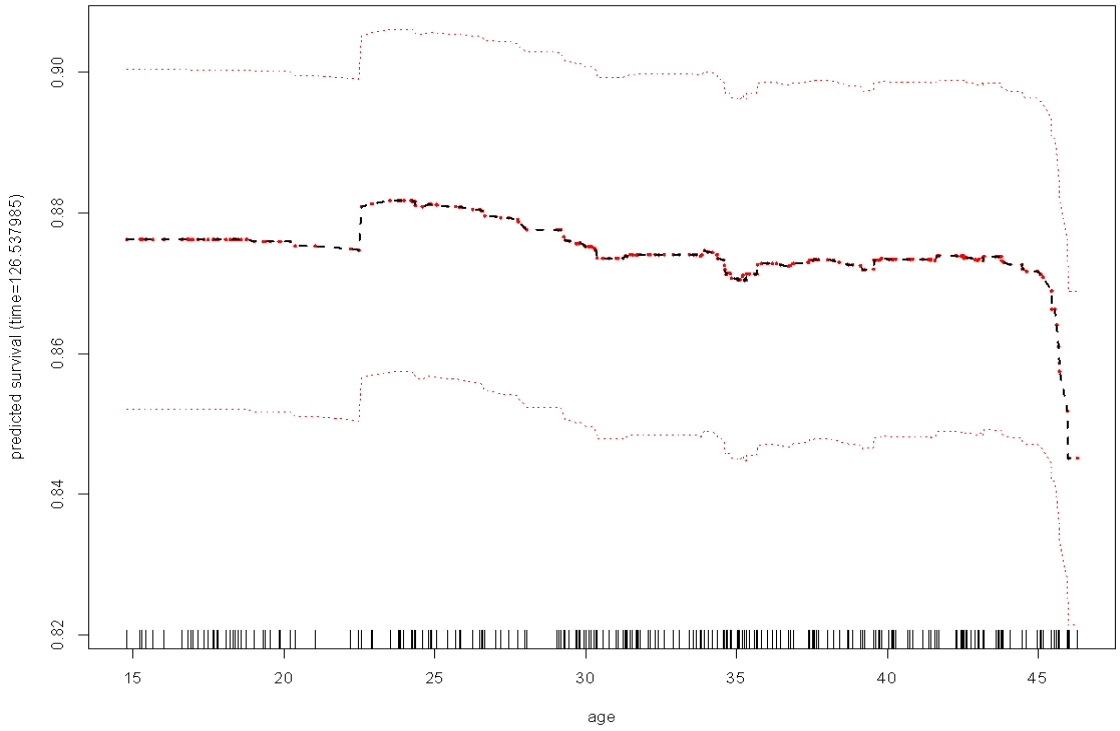

Figure 15. Marginal effect of age on predicted survival

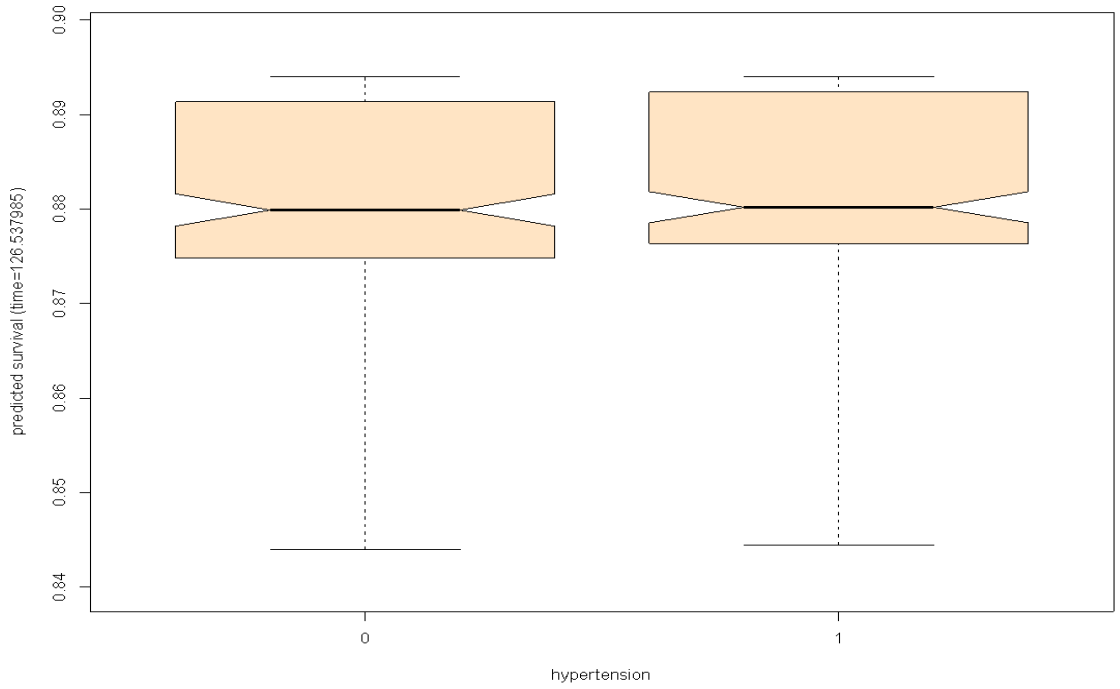

Figure 16. Marginal effect of hypertension on predicted survival

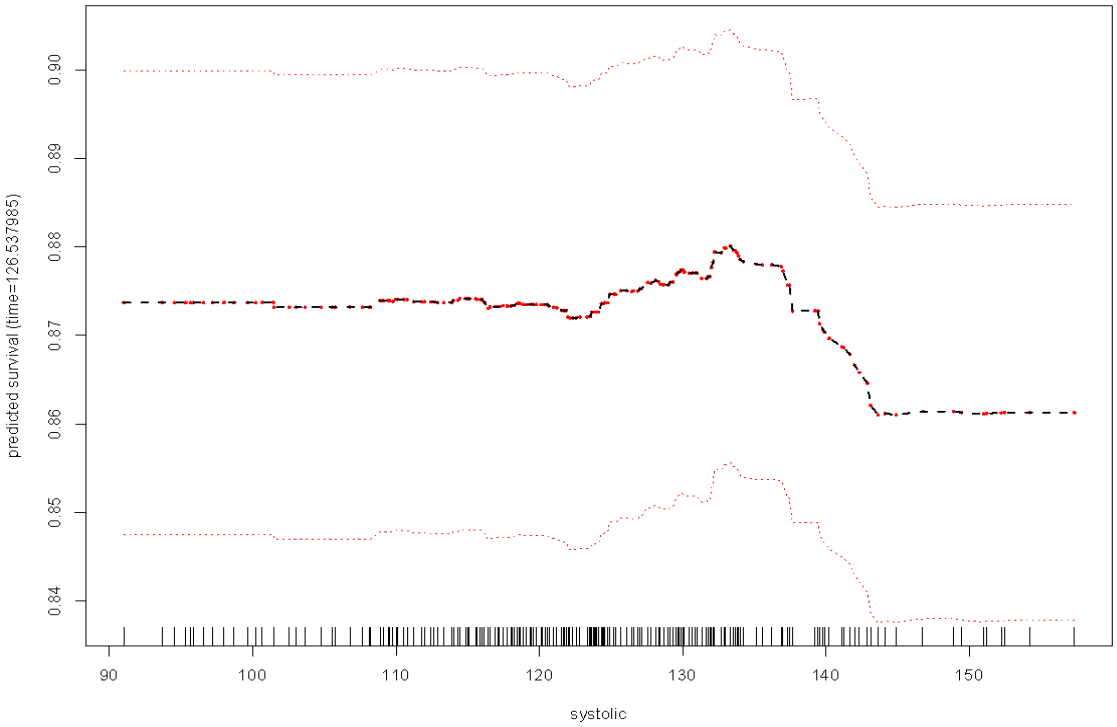

Figure 17. Marginal effect of systolic BP on predicted survival

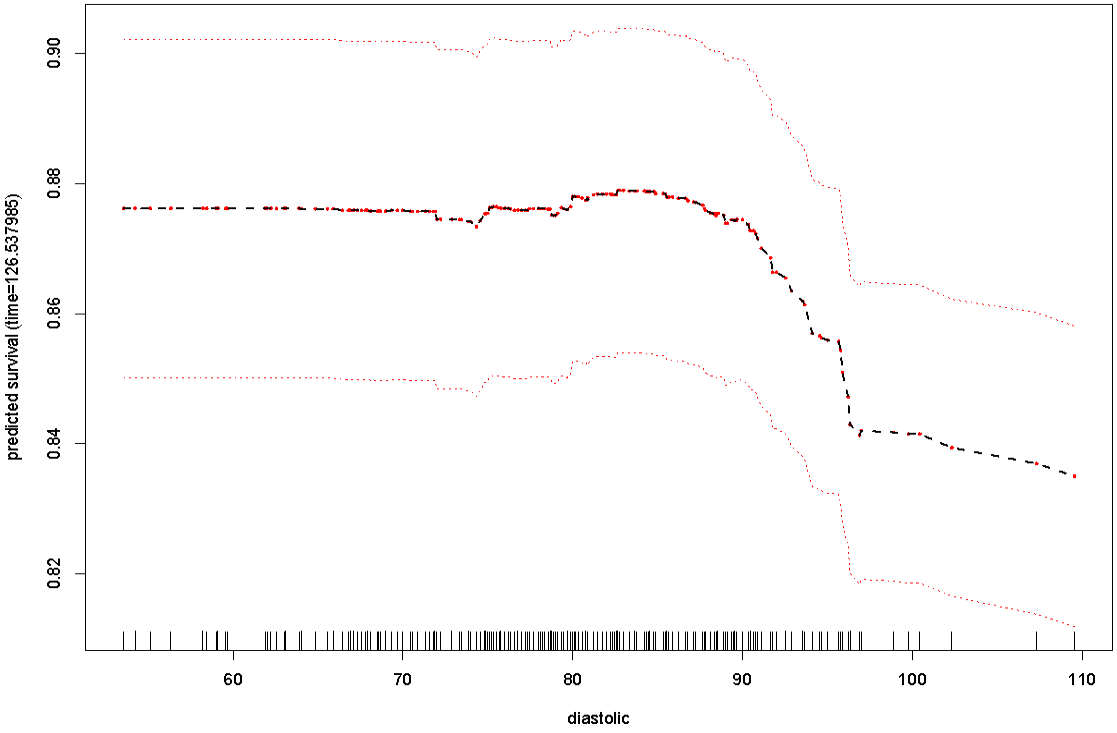

Figure 18. Marginal effect of diastolic BP on predicted survival

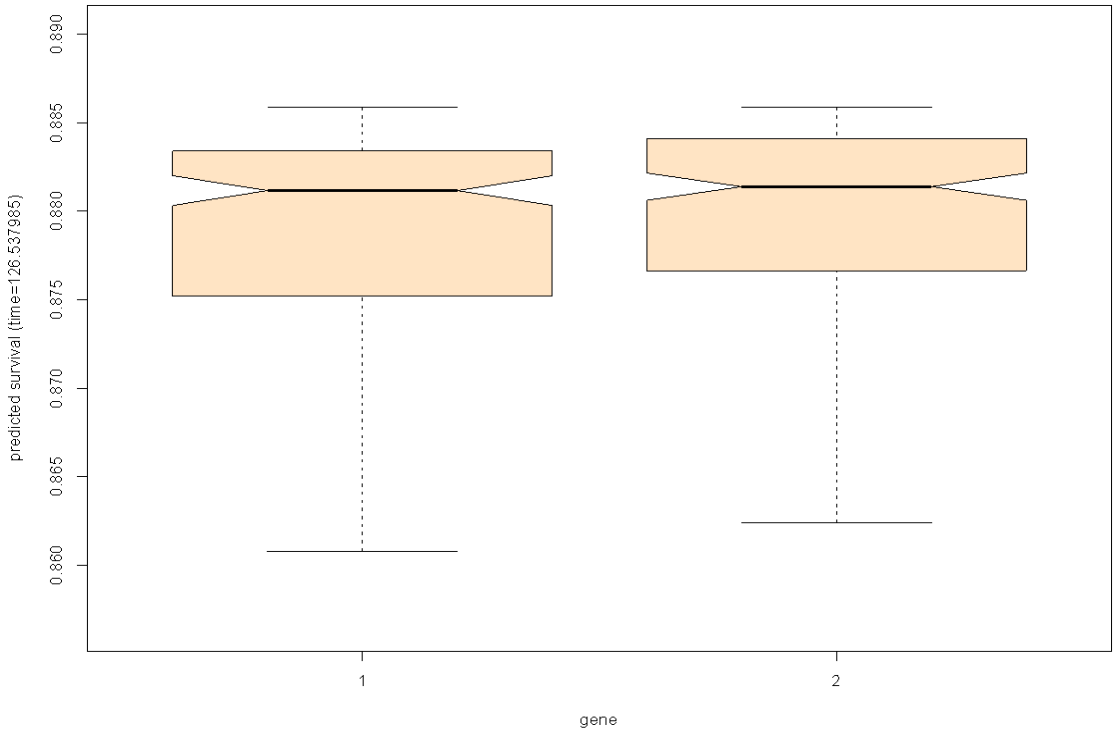

Figure 19. Marginal effect of genotype on predicted survival

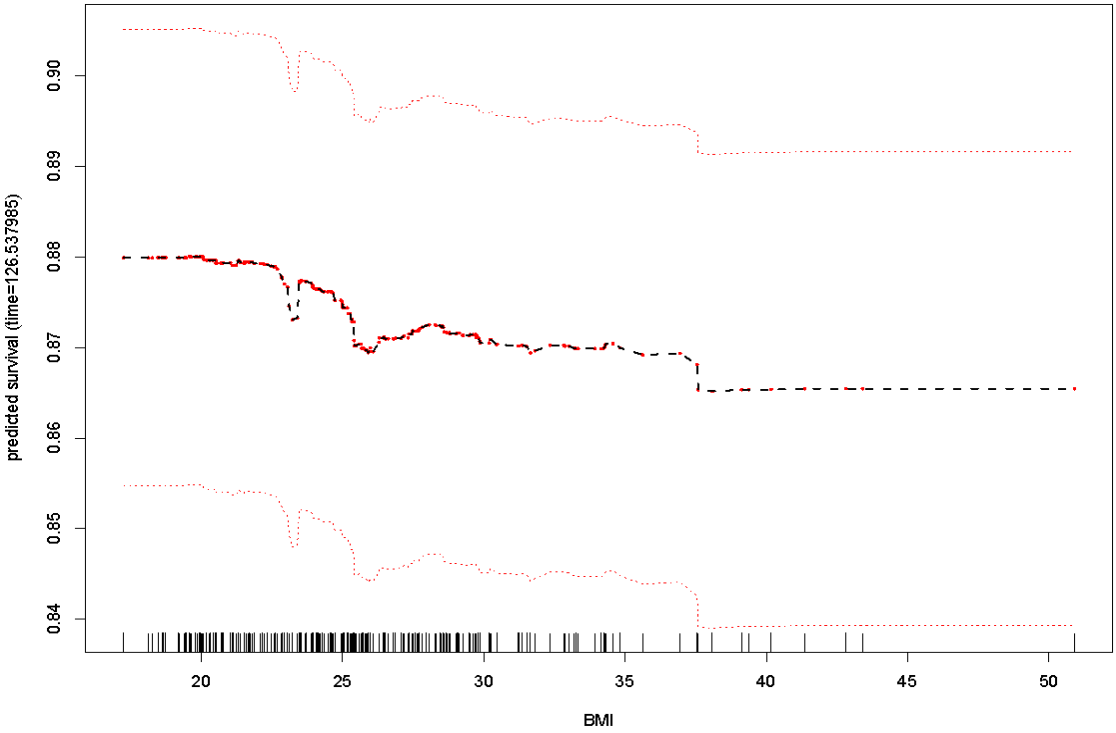

Figure 20. Marginal effect of BMI on predicted survival

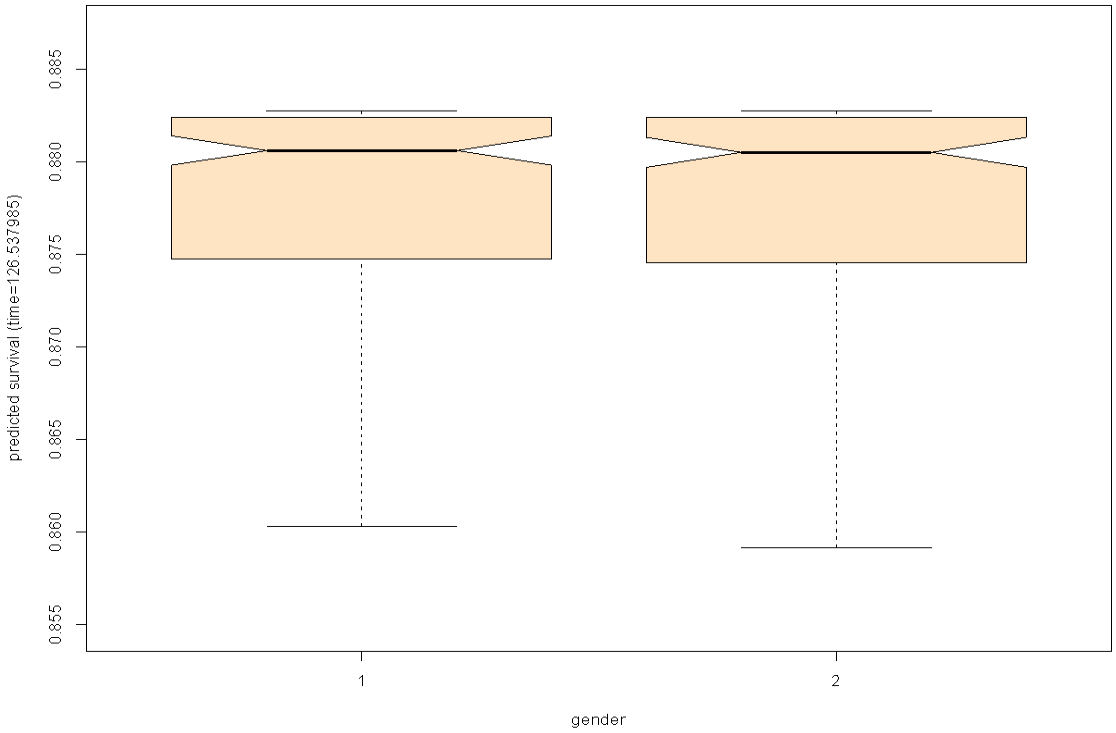

Figure 21. Marginal effect of gender on predicted survival

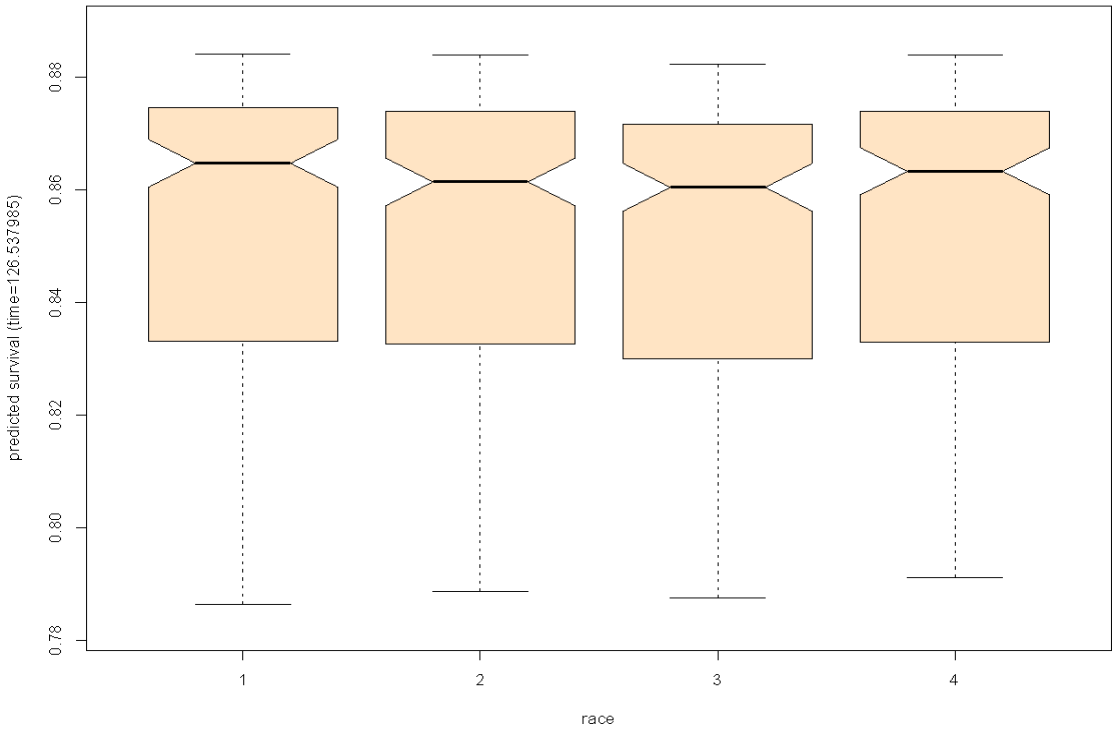

Figure 22. Marginal effect of race on predicted survival

Appendix B: Stata Code

```
/* Combining Stata Datasets */

clear

* Use ESRD dataset
use "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/Original Datasets/CRISP_ESRD_data.dta"

* Only keep relevant variables
keep pkdid basedate death deathdt esrd esrddt

* Create variables for the time, in months, until death or ESRD
gen death_month = ((deathdt - basedate) / 365.25) * 12
gen esrd_month = ((esrddt - basedate) / 365.25) * 12

* Save dataset
save "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/CRISP_ESRD_data.dta", replace

* Use Demographics dataset
```

```
use "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/Original Datasets/1_DEMOGRAPHICS_161116.dta"

* Only keep baseline data
keep if vis == 0

* Only keep relevant variables
keep pkdid age gender race4

* Rename race4 variable
rename race4 race

* Save dataset
save "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/1_DEMOGRAPHICS_161116.dta", replace

* Use Health dataset
use "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/Original Datasets/2_HEALTH_161116.dta"

* Only keep baseline data
keep if vis == 0
```

```
* Only keep relevant variables
keep pkdid bmi_c hypertensi~n diastol systol

* Rename variable
rename bmi_c BMI
rename hypertensi~n hypertension
rename diastol diastolic
rename systol systolic

* Save dataset
save "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/2_HEALTH_161116.dta", replace

* Use Imaging dataset
use "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/Original Datasets/3_IMAGING_161116.dta"

* Only keep baseline data
keep if vis == 0

* Only keep relevant variables
keep pkdid ckd_epi httkv tcv
```

```
* Rename variables
rename tcv TCV
rename ckd_epi GFR

* Save dataset
save "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/3_IMAGING_161116.dta", replace

* Use Cyst Growth dataset
use "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/Original Datasets/7_BIOMARKERS_161116.dta"

* Only keep baseline data
keep if vis == 0

* Only keep relevant variables
keep pkdid genetype

* Code the genetypes
gen gene = 2
replace gene = 1 if genetype == "PKD1"
drop genetype
```

```
* Save dataset
save "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/7_BIOMARKERS_161116.dta", replace

* Find the most recent follow up date
use "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/Original Datasets/1_DEMOGRAPHICS_161116.dta"

* Only keep most recent visit
collapse (last) visc, by(pkdid)

* Generate new variable of last follow up, in months
gen follow_up_months = visc*12
drop visc

* Save dataset
save "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/1_DEMOGRAPHICS_161116_followup.dta", replace

* Merge datasets
use "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/CRISP_ESRD_data.dta"
```

```
merge 1:1 pkdid using "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/1_DEMOGRAPHICS_161116.dta"
drop _merge

merge 1:1 pkdid using "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/2_HEALTH_161116.dta"
drop _merge

merge 1:1 pkdid using "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/3_IMAGING_161116.dta"
drop _merge

merge 1:1 pkdid using "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/1_DEMOGRAPHICS_161116_followup.dta"
drop _merge

merge 1:1 pkdid using "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/7_BIOMARKERS_161116.dta"
drop _merge


* If experienced ESRD, make ESRD date the last follow up date
replace follow_up_month = esrd_month if esrd_month != .
```

```
* Drop observation with no data
drop if pkdid == 403328
drop if pkdid == 203328
drop if follow_up_month == 0


* Convert gender to string variable
tostring gender, g (gender_2)
drop gender
rename gender_2 gender
* Male = 1, Female = 2

* Convert hypertension to string variable
tostring hypertension, g (hypertension_2)
drop hypertension
rename hypertension_2 hypertension
* Yes = 1, No = 2

* Convert race to string variable
tostring race, g (race_2)
drop race
rename race_2 race
```

* Caucasian = 1, African American = 2, Hispanic = 3, Asian = 4

** Descriptive Statistics

* Categorical variables

tab death

tab esrd

tab gender

tab race

tab hypertension

tab gene

tab death esrd

tab esrd esrd

tab gender esrd

tab race esrd

tab hypertension esrd

tab gene esrd

* Continuous variables

tabstat age diastolic systolic BMI GFR httkv TCV follow_up_months, statistics(count mean sd min p25 med p75 max) c(s)

```
* Save dataset
save "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/Combine_Dataset.dta", replace

* Export to CSV
export delimited using "/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/Combined Dataset.csv", replace
```

Appendix C: R Code

```
#-----------------------------------------------------------------------------------------------#

# Load all required packages
library(survival)
library(survminer)
library(plyr)
library(dplyr)
library(rpart)
library(rpart.plot)
library(party)
library(partykit)
library(rattle)
library(RColorBrewer)
library(data.table)
library(randomForestSRC)
library(ranger)
library(rms)
library(My.stepwise)
library(ggplot2)
```

```
# Read in and attach data
CRISP <- read.csv("/Volumes/Landsittel Group/CRISP/CRISP Projects/Alex/Thesis/New Datasets/Combined Dataset.csv",
                  header = TRUE)
attach(CRISP)

summary(CRISP)
summary(follow_up_months)

#-----------------------------------------------------------------------------------------------#

# Descriptive Statistics
par(mfrow = c(4,2), mar = c(2,4,0.5,0.5))
boxplot(age ~ esrd, names = c("No ESRD", "ESRD"), col = c("gray", "gray"), ylab = "Age", cex.label = 1.5)
boxplot(BMI ~ esrd, names = c("No ESRD", "ESRD"), col = c("gray", "gray"), ylab = "BMI")
boxplot(diastolic ~ esrd, names = c("No ESRD", "ESRD"), col = c("gray", "gray"), ylab = "Diastolic BP")
boxplot(systolic ~ esrd, names = c("No ESRD", "ESRD"), col = c("gray", "gray"), ylab = "Systolic BP")
boxplot(httkv ~ esrd, names = c("No ESRD", "ESRD"), col = c("gray", "gray"), ylab = "htTKV")
boxplot(TCV ~ esrd, names = c("No ESRD", "ESRD"), col = c("gray", "gray"), ylab = "TCV")
boxplot(GFR ~ esrd, names = c("No ESRD", "ESRD"), col = c("gray", "gray"), ylab = "GFR")
```

```
boxplot(follow_up_months ~ esrd, names = c("No ESRD", "ESRD"), col = c("gray", "gray"), ylab
= "Follow-up")

#-----------------------------------------------------------------------------------------------#



# Create survival object
survival_object <- Surv(time = follow_up_months, event = esrd)

# Use survival object to fit a Kaplan Meier estimate
survival_fit <- survfit(survival_object ~ 1)

# Plot the KM curve
plot(survival_fit, lwd=1, ylab="Survival Function", xlab="Months", xlim = c(0, 200))

# Create survival tree as a function of the the variabes listed below
survival_tree = rpart(formula = Surv(follow_up_months, event = esrd) ~ age + gender + race +
BMI +
        diastolic + systolic + hypertension + GFR + httkv + TCV + gene, data = CRISP,
        method = 'exp', control=rpart.control(minsplit = 20, minbucket = 10, maxdepth = 6))

# General plot of survival tree
plot(survival_tree, uniform = TRUE, margin = 0.1, main = "Survival Tree");
```

```
text(survival_tree, use.n = TRUE, cex = 0.9)

# Plot that shows proportions and frequencies
fancyRpartPlot(survival_tree)

# Print a list of all nodes, their splits, deviance, n, and y values
print(survival_tree)
summary(survival_tree)

# Find variable importance for single tree
rpart:::importance(survival_tree)
plot(rpart:::importance(survival_tree))

# Fit and plot easier to understand survival tree
(survival_tree_2 <- as.party(survival_tree))
plot(survival_tree_2, uniform = TRUE, cex = 0.5)

printcp(survival_tree)
plotcp(survival_tree)



#-------------------------------------------------------------------------------------------#
```

```
# Cox Proportional Hazards Model


# Cumulative Hazard: all patients
c_hazard_all <- -log(survival_fit$surv)
c_hazard_all <- c(c_hazard_all, tail(c_hazard_all, 1))


plot(c(survival_fit$time, 200),
     c_hazard_all,
     xlab = "Time (months)",
     ylab = "Cumulative Hazard",
     type = "s")


# Hazard model


# Undadjusted for each variable


cox_model_age          <- coxph(Surv(follow_up_months, event = esrd) ~ age, data = CRISP)
cox_model_age
cox_model_gender       <- coxph(Surv(follow_up_months, event = esrd) ~ gender, data = CRISP)
cox_model_gender
cox_model_race         <- coxph(Surv(follow_up_months, event = esrd) ~ race, data = CRISP)
cox_model_race
```

```
cox_model_BMI          <- coxph(Surv(follow_up_months, event = esrd) ~ BMI, data = CRISP)
cox_model_BMI
cox_model_diastolic    <- coxph(Surv(follow_up_months, event = esrd) ~ diastolic, data = CRISP)
cox_model_diastolic
cox_model_systolic     <- coxph(Surv(follow_up_months, event = esrd) ~ systolic, data = CRISP)
cox_model_systolic
cox_model_hypertension <- coxph(Surv(follow_up_months, event = esrd) ~ hypertension, data =
CRISP)
cox_model_hypertension
cox_model_GFR          <- coxph(Surv(follow_up_months, event = esrd) ~ GFR, data = CRISP)
cox_model_GFR
cox_model_httkc        <- coxph(Surv(follow_up_months, event = esrd) ~ httkv, data = CRISP)
cox_model_httkc
cox_model_TCV          <- coxph(Surv(follow_up_months, event = esrd) ~ TCV, data = CRISP)
cox_model_TCV
cox_model_gene         <- coxph(Surv(follow_up_months, event = esrd) ~ gene, data = CRISP)
cox_model_gene


cox_model_full <- coxph(Surv(follow_up_months, event = esrd) ~ age + gender + race + BMI +
                  diastolic + systolic + hypertension + GFR + httkv + gene, data = CRISP)
cox_model_full
```

```
cox_model_full_2 <- coxph(Surv(follow_up_months, event = esrd) ~ age + gender + race + BMI
+
                    diastolic + systolic + hypertension + GFR + TCV + gene, data = CRISP)
cox_model_full_2


test.cox <- cox.zph(cox_model_full)
test.cox
print(test.cox)
par(mfrow = c(2,2))
plot(test.cox)


test.cox.2 <- ggcoxzph(test.cox)
test.cox.2
plot(test.cox.2)


# Plot of residuals
res <- residuals(cox_model_full, type = "deviance")
plot(res)


#-------------------------------------------------------------------------------------------------#


node.2 = as.factor(survival_tree$where)
print(node.2)
```

```
CRISP_3 <- cbind(node.2, CRISP)
attach(CRISP_3)

# Create a list of colors
col = c("blue", "green3", "red")

# Plot Kaplan-Meier curves by class
plot(survfit(Surv(follow_up_months, event = esrd) ~ CRISP_3$node.2),
     ylab = "Survival Function", xlab = "Months", col = col)
legend("bottomleft", fill=col, inset=.02, col = col, cex = 1.0, pt.cex = 1.5, ncol = 1,
       title = "Class", legend=c("I", "II", "III"))

survdiff(Surv(time = follow_up_months, event = esrd) ~ node.2, data = CRISP_3)

# Pairwise differences
pairwise.other <- pairwise_survdiff(Surv(time = follow_up_months, event = esrd) ~ node.2, data
= CRISP_3)
pairwise.other

#-----------------------------------------------------------------------------------------------#

# Create variable for node, and merge with CRISP data
node = as.factor(survival_tree$where)
```

```
print(node)
CRISP_2 <- cbind(node, CRISP)
attach(CRISP_2)

# Create a list of colors
col = c("blue", "green3", "black", "red", "orange", "gray60", "brown", "purple")

# Plot Kaplan-Meier curves by node
plot(survfit(Surv(follow_up_months, event = esrd) ~ CRISP_2$node),
     ylab = "Survival Function", xlab = "Months", col = col)
legend("bottomleft", fill=col, inset=.02, col = col, cex = 1.2, pt.cex = 1.5, ncol = 2,
       title = "Node", legend=c("2", "5", "8", "9", "11", "13", "14", "15"))

# Recode nodes into classes
CRISP_2$class <- ifelse(node == 2,  1,
                 ifelse(node == 5,  2,
                 ifelse(node == 8,  3,
                 ifelse(node == 9,  4,
                 ifelse(node == 11, 5,
                 ifelse(node == 13, 6,
                 ifelse(node == 14, 7,
                 ifelse(node == 15, 8, 0)))))))
```

```
table(CRISP_2$class)

# Log-rank test between classes
survdiff(Surv(time = follow_up_months, event = esrd) ~ class, data = CRISP_2)

# Pairwise differences
pairwise <- pairwise_survdiff(Surv(time = follow_up_months, event = esrd) ~ class, data = CRISP_2)
pairwise

# Test
print(pairwise$p.value[[9]])

# Symbolic number coding
symnum(pairwise$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
      symbols = c("****", "***", "**", "*", "+", " "),
      abbr.colnames = FALSE, na = "")

# Create a list of colors
col = c("blue", "green3", "black", "red", "orange", "gray60", "brown", "purple")

# Plot Kaplan-Meier curves by class
plot(survfit(Surv(follow_up_months, event = esrd) ~ CRISP_2$class),
```

```
    main = "Kaplan-Meier Estimates by Class", ylab = "Survival Function", xlab = "Months", col
= col)
legend("bottomleft", fill=col, inset=.02, col = col, cex = 1.4, pt.cex = 0.7, ncol = 2,
       title = "Node", legend=c("1", "2", "3", "4", "5", "6", "7", "8"))

#-------------------------------------------------------------------------------------------#

### Consolidating Nodes ###

# Order by median survival time
ddply(CRISP_2,~class, summarise, median=median(follow_up_months))
ddply(CRISP_2,~class, summarise, count=length(class))

## Combination 1: Classes 3 and 5
CRISP_2$class.1 <- CRISP_2$class
CRISP_2$class.1[CRISP_2$class.1==5] <- 3

# Test
table(CRISP_2$class.1)

# Pairwise differences
pairwise.1 <- pairwise_survdiff(Surv(time = follow_up_months, event = esrd) ~ class.1, data =
CRISP_2)
```

```
pairwise.1

# Symbolic number coding
symnum(pairwise.1$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
      symbols = c("****", "***", "**", "*", "+", " "),
      abbr.colnames = FALSE, na = "")

# Order by median survival time
ddply(CRISP_2,~class.1, summarise, median=median(follow_up_months))

## Combination 2: Classes 1 and 3
CRISP_2$class.2 <- CRISP_2$class.1
CRISP_2$class.2[CRISP_2$class.2==3] <- 1

# Test
table(CRISP_2$class.2)

# Pairwise differences
pairwise.2 <- pairwise_survdiff(Surv(time = follow_up_months, event = esrd) ~ class.2, data =
CRISP_2)
pairwise.2

# Symbolic number coding
```

```
symnum(pairwise.2$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
      symbols = c("****", "***", "**", "*", "+", " "),
      abbr.colnames = FALSE, na = "")

# Order by median survival time
ddply(CRISP_2,~class.2, summarise, median=median(follow_up_months))

## Combination 3: Classes 1 and 2
CRISP_2$class.3 <- CRISP_2$class.2
CRISP_2$class.3[CRISP_2$class.3==2] <- 1

# Test
table(CRISP_2$class.3)

# Pairwise differences
pairwise.3 <- pairwise_survdiff(Surv(time = follow_up_months, event = esrd) ~ class.3, data =
CRISP_2)
pairwise.3

# Symbolic number coding
symnum(pairwise.3$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
      symbols = c("****", "***", "**", "*", "+", " "),
      abbr.colnames = FALSE, na = "")
```

```
# Order by median survival time
ddply(CRISP_2,~class.3, summarise, median=median(follow_up_months))

## Combination 4: Classes 4 and 6
CRISP_2$class.4 <- CRISP_2$class.3
CRISP_2$class.4[CRISP_2$class.4==6] <- 4

# Test
table(CRISP_2$class.4)

# Pairwise differences
pairwise.4 <- pairwise_survdiff(Surv(time = follow_up_months, event = esrd) ~ class.4, data =
CRISP_2)
pairwise.4

# Symbolic number coding
symnum(pairwise.4$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
      symbols = c("****", "***", "**", "*", "+", " "),
      abbr.colnames = FALSE, na = "")

## Combination 5: Classes 4 and 7
CRISP_2$class.5 <- CRISP_2$class.4
```

```
CRISP_2$class.5[CRISP_2$class.5==7] <- 4


# Test
table(CRISP_2$class.5)


# Pairwise differences
pairwise.5 <- pairwise_survdiff(Surv(time = follow_up_months, event = esrd) ~ class.5, data =
CRISP_2)
pairwise.5


# Symbolic number coding
symnum(pairwise.5$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
      symbols = c("****", "***", "**", "*", "+", " "),
      abbr.colnames = FALSE, na = "")



# Create final class variable
CRISP_2$class.final <- CRISP_2$class.5
CRISP_2$class.final[CRISP_2$class.final==4] <- 2
CRISP_2$class.final[CRISP_2$class.final==8] <- 3


# Log-rank test between classes
survdiff(Surv(time = follow_up_months, event = esrd) ~ class.final, data = CRISP_2)
```

```
# Summary of each class
ddply(CRISP_2,~class.final, summarise, median=median(follow_up_months))
ddply(CRISP_2,~class.final, summarise, count=length(class.final))
table(CRISP_2$class.final, esrd)
table(CRISP_2$node, esrd)



# Create a list of colors
col = c("blue", "green3", "red")

# Plot Kaplan-Meier curves by class
plot(survfit(Surv(follow_up_months, event = esrd) ~ CRISP_2$class.final),
     ylab = "Survival Function", xlab = "Months", col = col)
legend("bottomleft", fill=col, inset=.02, col = col, cex = 1.4, pt.cex = 1.5, ncol = 1,
       title = "Class", legend=c("I", "II", "III"))

#-----------------------------------------------------------------------------------------#

# Create a list of colors
col = c("blue", "blue", "blue", "green3", "blue", "green3", "green3", "red")

# Plot Kaplan-Meier curves by class
```

```
plot(survfit(Surv(follow_up_months, event = esrd) ~ CRISP_2$class),
     ylab = "Survival Function", xlab = "Months", col = col)
legend("bottomleft", fill=col, inset=.02, col = col, cex = 1.0, pt.cex = 1.5, ncol = 1,
       title = "Class", legend=c("I", "II", "III"))

#-----------------------------------------------------------------------------------------#

# Random Survival Forests

# Create a random survival forest object
forest <- rfsrc(Surv(follow_up_months, esrd) ~ age + gender + race + BMI +
                diastolic + systolic + hypertension + GFR + httkv + TCV + gene,
                data = CRISP,
                ensemble = "oob",
                importance = TRUE,
                block.size = 1,
                ntree = 500)

# Display results
print.rfsrc(forest)
print(forest)
summary(forest)
plot(forest, plots.one.page = FALSE)
```

```
# Plot survival
plot.survival(forest, plots.one.page = FALSE, cens.model = "km")

var.list <- c("age", "gene", "race", "gender", "httkv", "TCV",
              "GFR", "BMI", "hypertension", "diastolic", "systolic")

plot.variable(forest, xvar.names = var.list,
              surv.type = "surv", partial = T, oob = TRUE,
              show.plots = TRUE, plots.per.page = 1, granule = 5, sorted = TRUE,
              nvar, npts = 237, smooth.lines = F)
```

Bibliography

1. Disease, N.I.o.D.a.D.a.K. *What is Polycystic Kidney Disease?* January 2017; Available from: https://www.niddk.nih.gov/health-information/kidney-disease/polycystic-kidney-disease/what-is-pkd.
2. Halvorson, C.R., M.S. Bremmer, and S.C. Jacobs, *Polycystic kidney disease: inheritance, pathophysiology, prognosis, and treatment.* Int J Nephrol Renovasc Dis, 2010. **3**: p. 69-83.
3. Hateboer, N., et al., *Comparison of phenotypes of polycystic kidney disease types 1 and 2. European PKD1-PKD2 Study Group.* Lancet, 1999. **353**(9147): p. 103-7.
4. Patel, V., R. Chowdhury, and P. Igarashi, *Advances in the pathogenesis and treatment of polycystic kidney disease.* Curr Opin Nephrol Hypertens, 2009. **18**(2): p. 99-106.
5. Adams, M., *The Primary Cilium: An Orphan Organelle Finds a Home.* Nature Education, 2010. **3**(9).
6. Bennett, W.M., *Autosomal dominant polycystic kidney disease: 2009 update for internists.* Korean J Intern Med, 2009. **24**(3): p. 165-8.
7. Wüthrich, R.P., A.L. Serra, and A.D. Kistler, *Autosomal Dominant Polycystic Kidney Disease: New Treatment Options and How to Test Their Efficacy.* Kidney and Blood Pressure Research, 2009. **32**(5): p. 380-387.
8. Grantham, J.J., *Autosomal Dominant Polycystic Kidney Disease.* Annals of Transplatation, 2010. **14**(4): p. 5.
9. Torres, V.E., et al., *Tolvaptan in patients with autosomal dominant polycystic kidney disease.* N Engl J Med, 2012. **367**(25): p. 2407-18.
10. Fick, G.M., et al., *Causes of death in autosomal dominant polycystic kidney disease.* J Am Soc Nephrol, 1995. **5**(12): p. 2048-56.
11. System, U.S.R.D., *USRDS Annual Data Report.* 2016. **2**: p. 391.
12. US Department of Health and Human Services, C.f.D.C.a.P., *National Chronic Kidney Disease Fact Sheet.* 2017.
13. Repository, N.I.o.D.a.D.a.K.D.C., *Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP).*
14. Franz, K.A. and F.C. Reubi, *Rate of functional deterioration in polycystic kidney disease.* Kidney Int, 1983. **23**(3): p. 526-9.
15. Chapman, A.B., et al., *Renal structure in early autosomal-dominant polycystic kidney disease (ADPKD): The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) cohort.* Kidney Int, 2003. **64**(3): p. 1035-45.
16. Grantham, J.J., et al., *Volume progression in polycystic kidney disease.* N Engl J Med, 2006. **354**(20): p. 2122-30.
17. Harris, P.C., et al., *Cyst number but not the rate of cystic growth is associated with the mutated gene in autosomal dominant polycystic kidney disease.* J Am Soc Nephrol, 2006. **17**(11): p. 3013-9.
18. Chapman, A.B., et al., *Kidney volume and functional outcomes in autosomal dominant polycystic kidney disease.* Clin J Am Soc Nephrol, 2012. **7**(3): p. 479-86.

19. Peralta, C.A., et al., *Racial and ethnic differences in kidney function decline among persons without chronic kidney disease.* J Am Soc Nephrol, 2011. **22**(7): p. 1327-34.
20. Prakash, S. and A.M. O'Hare, *Interaction of aging and chronic kidney disease.* Semin Nephrol, 2009. **29**(5): p. 497-503.
21. Iseki, K., *Gender differences in chronic kidney disease.* Kidney Int, 2008. **74**(4): p. 415-7.
22. Dua, S., et al., *Body mass index relates to blood pressure among adults.* North American journal of medical sciences, 2014. **6**(2): p. 89-95.
23. Chang, T.-J., et al., *Relationship between body mass index and renal function deterioration among the Taiwanese chronic kidney disease population.* Scientific reports, 2018. **8**(1): p. 6908-6908.
24. Zaman, S.B., N. Hossain, and M. Rahman, *Associations between Body Mass Index and Chronic Kidney Disease in Type 2 Diabetes Mellitus Patients: Findings from the Northeast of Thailand.* Diabetes & metabolism journal, 2018. **42**(4): p. 330-337.
25. Bewick, V., L. Cheek, and J. Ball, *Statistics review 12: survival analysis.* Crit Care, 2004. **8**(5): p. 389-94.
26. Kartsonaki, C., *Survival analysis.* Diagnostic Histopathology, 2016. **22**(7): p. 263-270.
27. Strobl, C., J. Malley, and G. Tutz, *An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.* Psychol Methods, 2009. **14**(4): p. 323-48.
28. Gareth James, D.W.T.H.R.T., *An introduction to statistical learning : with applications in R*. 2013: New York : Springer, [2013] ©2013.
29. Wang, P., Y. Li, and C.K. Reddy, *Machine Learning for Survival Analysis: A Survey.* ACM Comput. Surv., 2019. **51**(6): p. 1-36.
30. Molinaro, A.M., S. Dudoit, and M.J. van der Laan, *Tree-based multivariate regression and density estimation with right-censored data.* Journal of Multivariate Analysis, 2004. **90**(1): p. 154-177.
31. Bou-Hamad, I., D. Larocque, and H. Ben-Ameur, *A review of survival trees.* Statist. Surv., 2011. **5**: p. 44-71.
32. Paravati, A.J., et al., *Radiotherapy and temozolomide for newly diagnosed glioblastoma and anaplastic astrocytoma: validation of Radiation Therapy Oncology Group-Recursive Partitioning Analysis in the IMRT and temozolomide era.* J Neurooncol, 2011. **104**(1): p. 339-49.
33. Ishwaran, H., et al., *Random survival forests.* Ann. Appl. Stat., 2008. **2**(3): p. 841-860.
34. Breiman, L., *Random Forests.* Mach. Learn., 2001. **45**(1): p. 5-32.