Predicting Outcomes in Lost-to-follow-up Subjects from a 15-year Observational Study of Autosomal Dominant Polycystic Kidney Disease

by

Anni Guo

MS, Southern Medical University, 2018

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

2020

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Anni Guo

It was defended on

April 20, 2020

and approved by

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

Jeanine M. Buchanich, PhD, Research Associate Professor of Biostatistics, Deputy Director of Center for Occupational Biostatistics and Epidemiology, Director of Biostatistics Consulting Laboratory, Graduate School of Public Health, University of Pittsburgh

Jenna C Carlson, PhD, Assistant Professor of Biostatistics, Graduate School of Public Health, University of Pittsburgh

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

Predicting Outcomes in Lost-to-follow-up Subjects from a 15-year Observational Study of Autosomal Dominant Polycystic Kidney Disease

Anni Guo, MS

University of Pittsburgh, 2020

Abstract

Autosomal dominant polycystic kidney disease (ADPKD) is a common chronic hereditary kidney disease, mainly characterized by kidney volume growth and cyst formation. Chronic Kidney Disease is the predictable result of ADPKD, which is usually defined as 5 stages (CKD, stage 1-5) from mild to severe by estimated Glomerular Filtration Rate (eGFR) values.

The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) is a study of ADPKD patients' kidney function decline. The participants were in different CKD stages and typically progress to worse CKD stages over time. CRISP was established, in large part, to describe the natural history of ADPKD. Given the necessary long-term follow-up, CRISP participants are often lost to follow-up (LTF).

In order to better use data from the LTF participants, this study focuses on predicting the CKD status at for the LTF participants in CRISP and assessing whether there is a difference between the LTF participants and the non-LTF participants. To predict the trajectory of eGFR, participants were grouped based on the Mayo imaging classification (MIC), which uses age and height-adjusted total kidney volume (htTKV) to estimate the rate of htTKV growth. Within each MIC, a different mixed model was fit to predict eGFR trajectory; the final status of each LTF participant was then estimated based on that trajectory. Bootstrapping was used to assess the variability of the predictions.

Results described the predicted CKD status and showed a minimal impact of variability on the prediction, allowing us to effectively predict the final outcome of LTF ADPKD patients. Further, the predicted outcomes of the LTF participants were consistent with the observed outcome of the non-LTF participants, which indicated the group of LTF participants was a random subset of the entire cohort.

The findings of the study are significant to public health. Because lack of follow-up will affect the effectiveness of the study, it is important to obtain the information of the LTF participants as much as possible. For ADPKD, an early understanding of the variability in patients with different disease risks could provide specific information for the development of disease, which is of great significance for proper prevention and treatment in the future.

Table of Contents

List of Tables

List of Figures

Preface

First of all, I would like to thank my thesis advisor, Dr. Douglas Landsittel for his constant encouragement and guidance in this process. He has walked me through all the stages of this thesis, offered me numerous valuable comments and suggestions with incomparable patience. At the same time, he also helped me improved my research and organization skills, which will continuously benefit me a lot in my future career life.

I would like to thank Dr. Ada Youk, Dr. Jeanine Buchanich, and Dr. Jenna Carlson for serving on my thesis committee. At the same time, they provided a lot of guidance and help for my thesis in Capstone class, without their consistent and illuminating instruction, this thesis could not have reached its present form. In addition to this, their warm encouragement also supported me through those stressful days.

I would like to thank my academic advisor Dr. Gong Tang, for his great support and instructions. It was he who pointed me out when I got into difficult problem. He cheered me up so that I got the motivation and courage to solve the following problems. His warm caring and guidance will benefit me for a whole lifetime.

I would also like to thank Avantika Srivastava of the Biomedical Statistics & Data Science Lab at the University of Pittsburgh for providing the data used for this analysis and providing meaningful analytical assistance, as well as the warm encouragement.

I would also like to thank all the professors and classmates in the Department of Biostatistics at the University of Pittsburgh. They have provided me with guidance and help for the past two years. The time I spent with everyone will be a wonderful experience in my life.

Finally, I would like to thank my family, especially my parents, Dongsheng Guo, Lei Shi, and my boyfriend, Ruopu Song, for their continuous encouragement and support. In the two years of studying for the Master's degree at the University of Pittsburgh, their love and companionship were a huge source of strength and motivation for me. Without them, I could not have done it.

1.0 Introduction

1.1 Autosomal Dominant Polycystic Kidney Disease (ADPKD)

Autosomal dominant polycystic kidney disease (ADPKD), is a type of hereditary kidney disease. As a common kidney disorder, ADPKD is estimated to have a diagnosed prevalence of 1:2000 and incidence of 1:3000-1:8000 on a global scale. In the United States, there are about 6000 new cases are diagnosed each year. The typical clinical manifestations of ADPKD are the increase in the size of the kidney and the formation of kidney cysts, which is also accompanied by many extrarenal complications, such as liver cysts, intracranial aneurysms, and Cardiac Valvular Disease (CVD). [1] Patients with ADPKD may experience the following symptoms: abdominal pain, haematuria, serious upper urinary tract infections (UTIs), kidney stones, and several other symptoms. Mild symptoms may interfere with the patient’s normal life, but severe illness may cause great pain to the patient.

ADPKD is a type of genetic condition where patients have a genetic mutation that accerates kidney volume growth and cyst formation. However, ADPKD rarely causes clinical symptoms and adverse affect kidney function later in life, e.g. around 30 to 60 years of age. Some of the variability in kidney function relates to the specific ADPKD mutation, which are PKD1, PKD2, or there may be no mutation detected (NMD). PKD1 accounts for nearly 85% of mutations. Patients with a PKD1 mutation usually have a greater number of cysts and thus an early onset of decline in renal function and presence of clinical symptoms. [2] [3] The degree of renal decline is often quantified by the condition of chronic kidney disease (CKD), where the kidney’s rate of filtration drops below normal function. As kidney function continues to decline, the patient may eventually reach End-Stage Kidney Disease (ESKD), where the kidneys fail to work and a transplant or dialysis is needed. [1]

ADPKD is one of the leading causes of ESKD; more than 50% of ADPKD patients eventually develop ESKD. Although there is no cure for ADPKD, but there are treatments for symptoms or diseases caused by this disease. Treatments include drugs to treat urinary tract infections or high blood pressure, and surgery to remove kidney stones.

1.2 Chronic Kidney Disease (CKD)

As described above, CKD is defined as a loss of kidney function over time (usually over decades). CKD is a predictable outcome for patients with the genetic condition of ADPKD . [4] [5] In addition to the symptoms associated with kidney function, CKD is associated with other clinical complications, including heart disease, high blood pressure, bone disease, and anemia. [4] [6] As described by the National Kidney Foundation (NKF), there are guidelines that divide CKD into five stages from mild to severe, in order to help physicians identify the levels of kidney disease. [7] [4] The stages of CKD are based on the estimated Glomerular Filtration Rate (eGFR) values. GFR is a commonly used measure of kidney function to quantify how well the kidneys filter blood. GFR can either be measured directly by iothalamate clearance or estimated using serum creatinine and certain personal characteristics. In our study, eGFR is calculated by the creatinine-based Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation, using the information of a person's age, gender, height, weight, race and creatinine levels:

$$\mathrm{GFR}_{\mathrm{CKD-EPI}} = 141 \times min(\mathrm{Scr}/\kappa, 1)^{\alpha} \times max(\mathrm{Scr}/\kappa, 1)^{-1.209} \times 0.993^{\mathrm{Age}} \times 1.018\ [\text{If Gender} = \text{Female}] \times 1.159\ [\text{If Race} = \text{Black}].$$

In the above formula, Scr is serum creatinine, the values of κ and α are constants; κ = 0.7 for females and κ = 0.9 for males; α = –0.329 for females and α = –0.411 for males, min indicates the minimum of Scr/κ or 1, and max indicates the maximum of Scr/κ or 1. [8] [9]

Stages of CKD are displayed in Table 1 [10]:

Table 1 CKD Stage Description

Stage	GFR (ml/min)	Description	% of kidney function
1	≥ 90	Normal functioning kidney	>90%
2	[60, 90]	Mild decrease in kidney function	60%-90%
3a	[45, 60]	Mild-moderate decrease in kidney function	45%-59%
3b	[30, 45]	Moderate-severe decrease in kidney function	30%-44%
4	[15, 30]	Severe decrease in kidney function	15%-29%
5	≤ 15	Kidney failure-ESKD	<15%

CKD stage 1 represents completely normal kidney function and stage 5 represents a nearly complete loss of kidney function. CKD stages 3 (also sometimes broken into 3a and 3b.) is usually considered the cut-off for having CKD, as kidney function loss in stage begins to result in clinical symptoms. At stage 4, the individual has significant kidney damage. Development of CKD stage 5 often requires a transplant or dialysis. [11] [12] In addition to presence of ADPKD, a number of factors can affect the rate of GFR decline over time. [13]

1.3 Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP)

The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) is an ongoing observational cohort study of 241 ADPKD participants that seeks to characterize the natural history of ADPKD and measure and evaluate biomarkers for prognosis of CKD and ESKD. For some of those biomarkers, characteristics of kidney structure and function are captured using high-resolution magnetic resonance (MR) imaging, including kidney volume, renal blood flow, and number and volume of cysts. [14] Data on baseline covariates (e.g. age, sex, and PKD mutation) was collected in 2001, and participants have been followed longitudinally with initially annually (through 3 follow-up visits over the first 5 years of the study) and then approximately bi-annually (with 4 planned visits) for the subsequent 10 years.

1.4 Motivation

As a long-term cohort study, one of the main challenges in CRISP is obtaining complete data collection on subjects over the extended time of follow-up. Some of participants may withdraw from the study for various reasons. Reasons may be random, such as having moved away from the clinical site during the study, or non-random reasons, such as becoming too ill to continue. This means that we do not know the final result of the participants' CKD development, so that these participants are considered to be lost to follow-up (LTF). Lost to follow-up is very important in determining a study's validity because LTF participants may tend to have a different prognosis than those who complete the study.

The CRISP data used for the current analysis was closed in 2016, after 15 years of follow-up. The timepoint of a participant entered the cohort study will be defined as Year 0. We picked Year 12 as a cut-off point for lost to follow-up, because clinic visits are spread out during the study. So a participant that fully participates may still have their final visit a few years before Year 15. In contrast, if their last visit was before Year 12, we considered them LTF.

For the purposes of the current study, we on whether the participant reached ESKD or CKD Stage 5 (subsequently referred to as stage 5), and whether the predicted outcome of LTF partipants differs from those who continued past Year 12. There were 44 subjects, who did not reach stage 5 and did not have eGFR data past Year 12; they were defined as lost-to-follow-up (LTF).

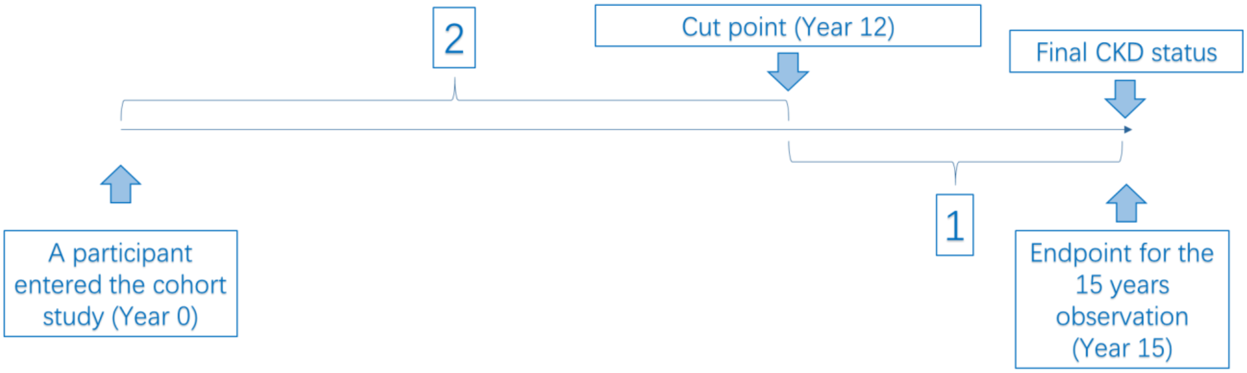

2
Cut point (Year 12)
Final CKD status
1
A participant entered the cohort study (Year 0)
Endpoint for the 15 years observation (Year 15)

Figure 1 Definition of Lost to follow-up Participants

(1). Do not have an eGFR measured after Year 12. (2) Without ESKD or CKD stage 5 during observation.

CRISP investigators assume that LTF participants are essentially a random subset. For instance, some of the participants have to move because of family reasons or other "random" events that cause them to become LTF. The question is whether the group of LTF participants is similar to those with adequate follow-up or observed end point outcomes, or whether the LTF group is a biased subset that looks much worse than those still under study. If we see that the predicted proportion of stage 5 is much higher in those LTF as compared to those with complete follow-up (greater than12 years or reach ESKD), this would indicate potential for a bias in who is LTF. In contrast, we may see the predicted proportion of 5 stage is much lower in the LTF since they necessarily exclude those with ESKD, and data capture of ESKD is generally considered to be near complete.

For the purposes of this current study, our first step was to predict the eGFR values, and the corresponding predicted probability of reaching stage 5 at Year 15 for the LFT participants in CRISP. Second, we assessed the variability of the predicted eGFR values using bootstrapping. Predicted eGFR values and the predicted proportion reaching CKD stage 5 (which is a eGFR < 15 mm/min) in the LTF participants were then compared to the observed results in those not LTF.

2.0 Methodology

2.1 Description of Data Source

The CRISP data set used for this analysis was closed in 2016, after 15 years of follow-up. After removing participants who were defined as ‘atypical’, there were 236 participants included in our study. 41 participants reached stage 5 and 39 participants had dialysis or kidney transplant. For the 156 remaining subjects, 112 had follow-up past year 12; the remaining 44 subjects were defined as lost-to-follow-up (LTF). Summary statistics are displayed for baseline variables to describe the two groups.

2.2 Simple Linear Regression

While linear regression is not directly used in this study, the basic method is presented as some background before introducing linear mixed models (which are used to predict eGFR at year 15 for the LTF group). Simple linear regression estimates the average response for values of a given explanatory variable. Other subject characteristics, such as treatment or other demographic characteristics) could also be included in the estimation (which would be a multiple regression model).

For ith observation x_i,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

In the above equation, y_i is the measurement for the dependent random variable Y, β_0 is the true intercept, β_1 is slope, x_i is the measurement for the independent random variable X, and ε_i is the random error term which is assumed to be independently identically distributed (i.i.d.): $\varepsilon_i \sim N(0, \sigma^2)$, and $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, i.i.d..

The method of ordinary least squares (OLS) is used to estimate the coefficients in the model with the estimator:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

In the above equation, $\bar{x}$ is the sample mean of the independent variable, $\bar{y}$ is the sample mean of the dependent variables.

As noted above, the simple (and multiple) linear regression model method assumes independent observations. However, CRISP data has longitudinal measurements of the key outcome of eGFR; for this scenario, observations are not independent, thus yielding invalid standard errors. In other words, for the longitudinal data in CRISP, we are measuring the same subject multiple times so that these observations on the same subject are correlated. If we ignore the dependence between the observations, estimation of covariate effects will be biased leading to incorrect inference.

2.3 Mixed Effect Model

A longitudinal mixed model contains both a within-subjects factor (time-random factor) and a between-subjects factor (fixed factors). The mixed effects model for longitudinal data is given by:

$$y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

The parameter μ is the overall average outcome (e.g. eGFR), for *i*th subject, *j*th level, and *k*th time point, α_j is the fixed effect (usually the demographical variables), β_k is the fixed effect of time (the time each measurement occurs), and $\alpha\beta_{jk}$ is the time by fixed effect interactions. The error term ε_{ijk} can be estimated by a block diagonal covariance matrix R, in which each block corresponds to a different subject. The structure of the blocks in R reflects the researchers' assumptions about the pattern of error correlations within subjects. [15]

In previous studies of eGFR [11], the data were analyzed with polynomial mixed effect models with linear and quadratic terms of the covariate of age. Previous results of the CRISP study showed that the quadratic term of age in the mixed effects model was statistically significant; adding a quadratic term for age is also consistent with the underlying biology, where GFR may stay steady or increase in early life and then decrease sharply later in life. The polynomial mixed effect model that was subsequently established in the previous CRISP study is expressed as:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \zeta_{0j} + \varepsilon_{ij}$$

In the above equation, y_{ij} is the eGFR value for subject j at time i, x_{ij} is age of subject j at time i, x_{ij}^2 is quadratic term of age of child j at time i ζ_{0j} is the specific intercept deviation of subject j, and ε_{ij} is the random residual error.

2.4 Mayo Classification

CRISP and other studies have begun using a newly developed approach to categorize ADPKD patients based on their rate of height-adjusted TKV (htTKV) growth over time. Because htTKV growth is highly predictable over time [16], the rate of kidney growth can be reliably estimated by using a single htTKV measurement and the subject's current age. Irazabel, et al. [17] suggested five subgroups (see Table 2) to characterize the subject's rate of kidney growth referred to as the Mayo Classification. [17]

Table 2 Mayo Classification Description

Subgroup	htTKV Growth Rate/Year (%)
A	<1.5%
B	1.5%–3%
C	3%–4.5%
D	4.5%–6%
E	>6%

Over the follow up period, most CRISP participants had 4-8 eGFR measurements. In past studies, based on plotting the eGFR trajectory over time, Yu et al. indicated that for ADPKD patients with different severity of CKD, the processes of loss of kidney function have different trajectories. Patterns of GFR decline have been found to be significantly different in these five subgroups of the Mayo classification, all showing non-linear decline trajectories. [11] The participants with Mayo classification type A (slow kidney volume growth) tend to have normal ranges of eGFR, and will hardly reach CKD stage 5. Participants in categories B and C also tend to remain relatively normal, although some may drop to CKD stage 5 based on their initial eGFR. Class D and E participants (the group with the fastest kidney volume growth) tend to have very rapid renal failure in their 40s or 50s and reach CKD stage 5.

Yu et al. [11] showed that if a separate mixed-effect polynomial model is fitted in each Mayo classification, we will obtain a well-fitting model for predicting eGFR decline. The studies have shown that, although there is substantially variable in GFR trajectories over time, the Mayo Classification can be used to predict whether subjects will likely reach CKD stage 5 over a given period of time. Therefore, we will use existing methods (including Mayo classification and mixed effect models, where age and age squares are used as predictors, which are also affected by the random intercept of each subject) based on all their data to obtain the predicted values of GFR for the 44 LFT patients after 15 years passed from the baseline time point that the patients entered the cohort study.

Five mixed effect models were established for each Mayo classification with the formula:

$$\text{eGFR} = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Age}^2$$

Where β_0 is intercept, β_1 is the coefficient of participants' age (after 15 years passed from the baseline time point), and β_2 is the coefficient of participants' age^2 (age: after 15 years passed from the baseline time point). [11] [17] The models explained the relationship between increasing age and decreasing GFR values.

Table 3 Coefficients for the Published Mixed Effect Regression Polynomial Model

Class	CKD-EPI eGFR		
	β_0	β_1	β_2
A	79.16	1.47	-0.03
B	72.89	2.30	-0.05
C	121.64	0.37	-0.04
D	111.44	1.74	-0.07
E	98.06	2.87	-0.11

2.5 Bootstrap Method

The Bootstrap method is a statistical resampling technique which is used to estimate the statistics of the population. The basic idea for Bootstrap is: for a sample of size *n*, we sample from the original data with replacement. The probability of each observation in the original data being drawn each time is $\frac{1}{n}$, and the new obtained samples are called Bootstrap samples. Then an estimated value $\hat{\theta}$ of the parameter θ can be obtained for each bootstrap sample.

Let random sample $X = \{x_1,\ x_2, \ldots, x_n\}$ be independent and identically distributed samples, $x_i \sim F(x), i = 1, 2, \ldots, n$. $R(X, F)$ is a function of X and F. To estimate the distribution characteristics of $R(X, F)$ based on the observed samples $X = \{x_1,\ x_2, \ldots, x_n\}$, we let $\theta = \theta(F)$ be a parameter of the overall distribution *F*, F_n is the empirical distribution function of the observation sample *X*, and $\hat{\theta} = \hat{\theta}(F_n)$ is an estimate of *θ*, the estimation error is:

$$R(X, F) = \hat{\theta}(F_n) - \theta(F) \triangleq T_n$$

The basic steps for calculating the distribution characteristics of $R(X, F)$ are summarized as follows:

1) Construct the empirical distribution function F_n according to the observation sample $X = \{x_1, x_2, \dots, x_n\}$;
2) Take samples from F_n called Bootstrap samples;
3) Calculate the corresponding Bootstrap statistic $R^*(X^*, F_n)$, which can be expressed as:

$$R^*(X^*, F_n) = \hat{\theta}(F_n{}^*) - \hat{\theta}(F_n) \triangleq R_n$$

Where $F_n{}^*$ is the empirical distribution function of the Bootstrap sample; R_n is the Bootstrap statistic of T_n;

4) Repeat process 2) and 3) *N* times to get *N* possible values of Bootstrap statistic $R^*(X^*, F_n)$;
5) Use the distribution of $R^*(X^*, F_n)$ to approximate the distribution of $R(X, F)$, that is, use the distribution of R_n to approximate the distribution of T_n, getting *N* possible values of the parameter $\theta(F)$.

After these steps, we can estimate the distribution of the parameter θ.

The confidence interval is a commonly used interval estimation method. The confidence level represents the frequency (ie, the ratio) of possible confidence intervals that contain the true values of unknown population parameter $\theta(F)$.

Confidence intervals provide a range of model properties. When predicting new data from the model, it shows the likelihood that the model ’s predictions will fall between these ranges. For the bootstrap method, we use a non-parametric method to obtain the 95% confidence interval of the bootstrap estimate statistics. This form of confidence interval does not make any assumptions about the functional form of the statistical distribution, which is often called an empirical confidence interval. [18]

To obtain the 95% confidence interval, first, we sort the bootstrap estimated statistics, and then select the values at the selected percentile for the confidence interval. In this case, the selected percentile is called alpha. We take a 95% confidence interval in this study, and the alpha should be 0.95. We will choose a lower limit of 2.5% and an upper limit of 97.5% in the statistical data of interest.

In this study, we calculated 1,000 bootstrap estimates of the eGFR values using the predicted eGFR at Year 15 from the mixed model corresponding that subject’s Mayo classification. A total of 1,000 bootstrap samples were used to assess variability of the prediction. More specifically, the 1,000 predictions for each subject were ordered and the bootstrap confidence interval was defined as the range from the lower limit of the 25th value (i.e. 2.5 percentile) to the upper limit of the 975th value (i.e. 97.5 percentile). The overall prediction estimate of eGFR for the given subject was defined as the mean prediction over all 1,000 bootstrap samples. A chi-squared test was used to evaluate the difference in distributions between outcomes between the predictions in the LTF participants group versus the observed results in the non-LTF participants group.

3.0 Results

3.1 Summarized Statistics

Table 4 Baseline and Follow-up Information of CRISP Data

Variables	Mean (sd)/Numbers
Baseline age	32.26 (± 8.80)
Baseline height-adjusted total kidney volume (ml/m)	615.71 (± 371.32)
Baseline estimated glomerular filtration rate (mm/min)	92.48 (±22.97)
Mayo classification	
A	15 (6.36%)
B	59 (25.00%)
C	69 (29.24%)
D	55 (23.30%)
E	38 (16.10%)
Last observed GFR value (mm/min)	65.98 (± 33.92)
<12 years of follow-up	68.75 (± 21.29)
≥12 years of follow-up	65.37 (± 36.16)

Table 4 shows the baseline demographics and follow-up information on key characteristics. The mean baseline age of the participants was 32 years old, with a mean baseline height-adjusted total kidney volume (htTKV) equal to 615.71 ml/m, and a mean baseline GFR equal to 92.48 mm/min. There were 15 (6.36%) participants defined as Mayo classification subgroup A, 59 (25%) participants defined as subgroup B, 69 (29.24%) participants defined as subgroup C, 55 (23.3%) participants defined as subgroup D, and 38 (16.1%) participants defined as subgroup E. The participants with follow-up data past year 12 had a mean last observed GFR equal to 65.37 mm/min, and participants with follow-up data did not past year 12 had a mean last observed GFR equal to 68.75 mm/min.

3.2 Predicted Results of Mixed Models

Based on the bootstrap predicted eGFR, we classified the LTF participants' status into different CKD stages according to Table 1.

Table 5 Observed and Predicted Outcomes

	Observed Outcomes		Predicted Outcomes**
Outcome Status	Not lost to follow-up (%) N=192	lost to follow-up (%) N=44	lost to follow-up (%) N=44
CKD Stage 1-4	112 (58.3%)	44 (100%)	37 (84.10%)
CKD Stage 5	41 (21.35%)	0	7 (15.90%)
Transplant or Dialysis	39 (20.31%)	0	N/A*

*By definition, transplant and dialysis cases were not lost-to-follow-up (since follow-up ended with those endpoints).

**Using the linear mixed model prediction stratified by Mayo classification.

The observed CKD status as determined by the last eGFR measurement are shown in Table 5 for the 192 non-LTF participants and the 44 LTF participants. For the non-LTF participants, 112 out of 192 (58.3%) were distributed in CKD stage 1 to CKD stage 4, and 41 out of 192 (21.35%) were distributed in CKD stage 5, the remaining 39 (20.31%) participants had dialysis or had kidney transplants. By definition, participants that had transplant and dialysis were not lost-to-follow-up and placed into CKD stage 5. For all the 44 LTF participants, by the definition of lost to follow-up, none of them were observed to reach stage 5 (which we defined above to also include transplant or dialysis). Table 5 also includes the predicted CKD status at Year 15 for the LTF participants. Thirty-three out of 44 (75%) of the LTF participants were predicted to enter CKD stage 1-4, and 7 out of 44 (15.9%) of the LTF participants were predicted to entered CKD stage 5.

3.3 Comparison Between the LTF and non-LTF participants

The first Chi-square test was performed to test whether there is a difference between the observed outcome of CKD status for the non-LTF group and the last observed outcome for LTF group. By definition, the participants who had transplant or dialysis were not lost-to-follow-up and we included these participants into the CKD stage 5 category in the non-LTF group. The test showed a highly significant difference in the distributions of CKD stage ($p < 0.001$).

The second Chi-square test was performed to test whether there is a difference between the observed outcome of CKD status of the non-LTF group and the predicted outcome of CKD status of LTF group. This test also showed a highly significant statistically difference ($p = 0.003$).

In the last Chi-square test, we excluded the participants with transplant or dialysis in the non-LTF group, leaving 41 participants in the stage 5 category in the non-LTF group. This test result did not show significant evidence of a difference ($p = 0.20$).

3.4 Variability of the Prediction

In order to assess the variability of the predicted results, we obtained the 95% confidence interval of the predicted eGFR values from the results of the 1000 bootstrap samples for each LTF participant.

Table 6 Bootstrap Estimated Confidence Intervals for Each LTF Participants

PKDID	Predicted eGFR results	Estimated lower bound eGFR values (mm/min)	Estimated upper bound eGFR values (mm/min)	Confidence Interval Range
101585	40.69728381	32.80055409	48.32806574	15.527512
110080	45.52781638	38.45119642	52.5303288	14.079132
124300	-18.1575771	-38.6993049	0.831825318	39.53113
126133	53.10674238	47.06648205	59.10662857	12.040147
139126	23.83644708	14.34498292	32.75847472	18.413492
139486	58.52622913	53.19565422	64.02913262	10.833478
151030	65.43612562	26.5172416	101.4835439	74.966302
157925	47.11595925	32.87862303	60.41606549	27.537442
159106	-17.22714702	-34.9481604	-1.33548771	33.612673
160928	73.12228185	44.22419632	98.31974357	54.095547
161547	43.49753274	27.70950077	58.13826684	30.428766
170121	42.58607359	26.44361487	57.47134958	31.027735
174632	48.48095831	40.74091753	56.47280187	15.731884
193273	93.19847999	89.02268672	96.97687493	7.954188
194105	49.18083686	42.67824522	55.78861069	13.110365
195310	16.2052006	2.856998907	28.64369023	25.786691
223343	42.183622	34.4436392	49.69080254	15.247163
223534	41.6907207	34.42244963	48.64874748	14.226298
229428	49.64102834	43.14674492	56.16508073	13.018336
234650	70.46465327	65.98054413	75.00315204	9.022608
236202	-1.46459333	-15.5315141	11.55206201	27.083576
244111	80.55609091	76.56188008	84.45187722	7.889997
256171	57.64797842	47.73398763	67.17170105	19.437713
268455	37.61255369	28.21745705	46.81119075	18.593734
271662	35.01634132	25.27802386	44.23487087	18.956847
273214	52.33923618	46.2922519	58.39118179	12.09893
281977	46.83095289	40.18076431	53.3525621	13.171798
285601	0.22227941	-15.93869691	15.01838136	30.957078
293598	6.602780805	-8.043853523	20.25150795	28.295361
300911	30.88226106	20.60385737	40.65371875	20.049861

Table 6 Continued

313307	48.29998448	34.56533314	61.12978717	26.564454
320182	41.84832159	34.59907904	48.77222102	14.173142
333524	-60.5032608	-92.84873214	-30.73508699	62.113645
337315	64.88506202	60.01622656	69.79544425	9.779218
343097	23.27844031	13.69960926	32.26517626	18.565567
368973	87.33682813	83.82869707	90.8624546	7.033758
385151	46.38274098	39.67887176	52.92528579	13.246414
393936	46.17002127	39.460845	52.72241382	13.261569
394588	3.700605099	-9.633219476	15.90816275	25.541382
406726	96.90678149	92.89960229	101.0256293	8.126027
407648	43.38818101	27.55672426	58.05822761	30.501503
407841	57.84644868	48.01296056	67.34093867	19.327978
430543	40.07922576	32.05261557	47.74792584	15.69531
476972	69.12819449	62.81637155	75.83287343	13.016502

Table 6 shows the mean of the bootstrap predicted eGFR values, lower bound and upper bounds of the 95% confidence interval of the predicted eGFR values at 15-year follow-up for each LTF participant, and the range of the 95% confidence interval (CI range). Because eGFR = 15 mm/min was a cut-off point of the dichotomized results, CKD stage 1-4 and CKD stage 5, we were interested in the LTF participants who had a 95% confidence interval containing eGFR = 15 mm/min. We could see that there were just a few of the participants (n=6) had a confidence interval that contained eGFR = 15 mm/min, highlighted in yellow. Thirty-eight out of the 44 LTF participants had a 95% confidence interval that did not contain GFR value = 15 mm/min. As eGFR value decreases by 30 mm/min, the CKD stage deteriorates by one level, we observed those participants whose confidence interval range exceeds 30 (n=9) and found that there were only one of them had the 95% confidence interval including eGFR = 15 mm/min, the others all had either confidence interval upper bound smaller than 15 or lower bound larger than 15, clearly classified to CKD stage 5 or CKD stage 1-4

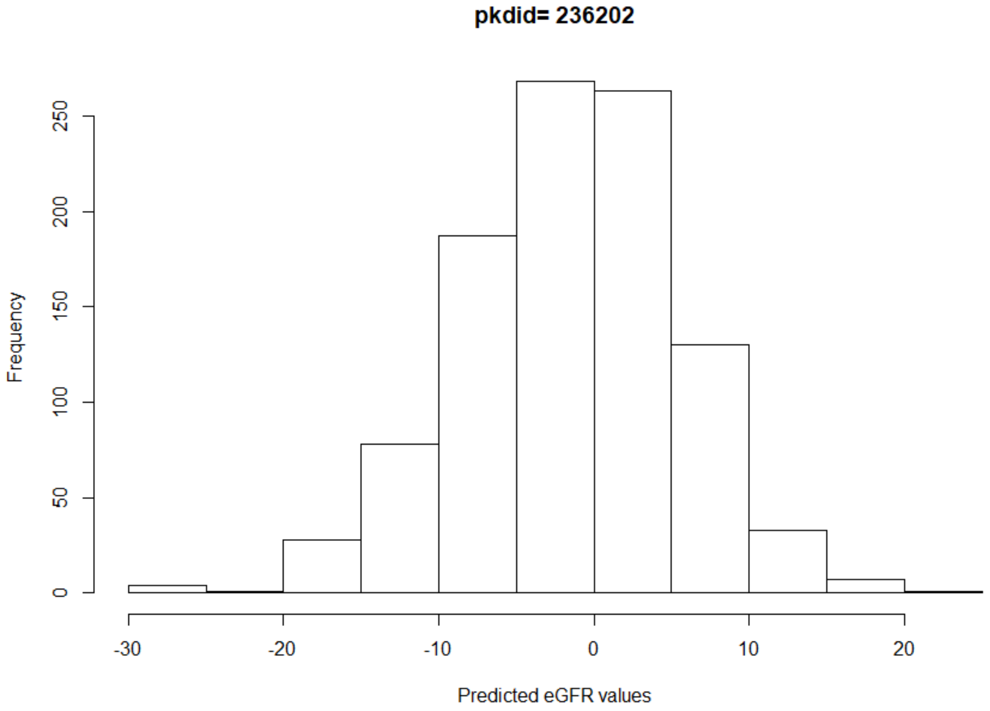


Figure 2 Histogram of Bootstrap Predicted Results, pkdid=236202

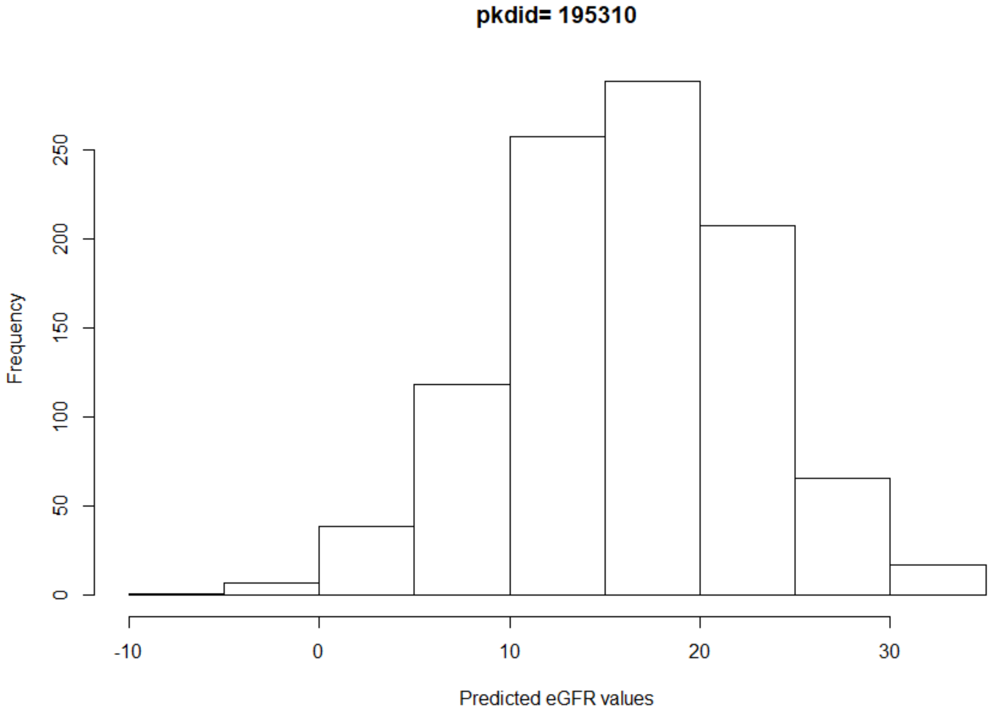


Figure 3 Histogram of Bootstrap Prediced Results, pkdid=195310

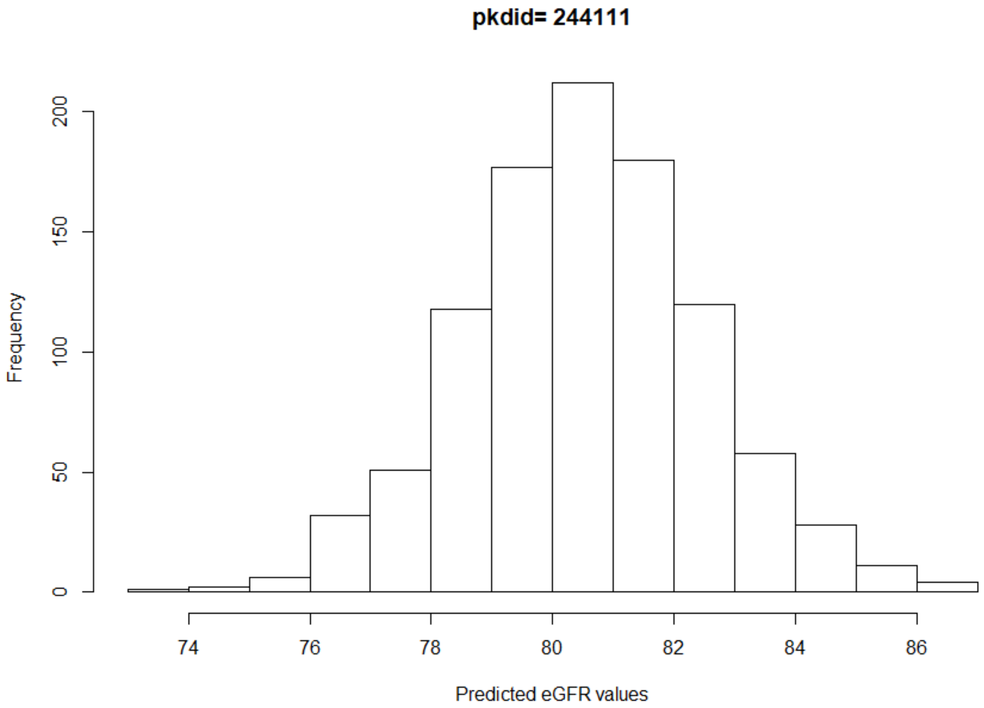


Figure 4 Histogram of Bootstrap Predicted Results, pkdid=244111

Figure 2, 3 and 4 showed the histograms of the bootstrap predicted results of certain LTF participants. The x-axis represented the bootstrap predicted eGFR values (mm/min) at Year 15 and the y-axis showed the number of times the predicted GFR value fell into a certain range out of 1000 bootstrap predictions. Also, eGFR = 15 mm/min was a cut-off point of the dichotomized results, CKD stage 1-4 and CKD stage 5. As shown in the figures, participant pkdid = 236202 had most of the predicted eGFR values smaller than 15 mm/min and only very few of the predicted results fell into the range of 10-20 mm/min, so that the participant was tent to be classified as CKD in stage 5. On the contrary, the participant pkdid = 244111 had all bootstrap predicted eGFR values larger than 15 mm/min, which indicated that the participants was tent to be classified as in CKD stage 1-4. Finally, the participant pkdid = 195310 had most of the bootstrap predicted eGFR values fell into range 10-20 mm/min, which showed that the predicted CKD stages of this participants contain both CKD stage 1-4 and CKD stage 5. Overall, these results showed that, while there is some sampling variability reflected in the results in predictions, that variability was low enough to yield estimates that were relatively stable in terms of whether a given subject was or was not predicted to reach stage 5.

4.0 Discussion

For the 44 LTF participants in the CRISP study, we used established mixed models, specific to a subject's Mayo classification (i.e. their rate of kidney volume increases) to predict the eGFR at Year 15. Then, the predicted proportion reaching stage 5 was then compared with the observed outcomes of non-LTF participants to evaluate whether there were differences between the two participant groups. Bootstrapping was also used to assess sampling variability and calculate confidence intervals for the eGFR predictions.

In previous studies, the Mayo classification was found to be an effective risk stratification tool that is the new standard in predicting kidney outcomes in high risk patients. However, the Mayo classification did not give the researchers the prediction specifically. Yu, et al., fitted separate mixed models within each Mayo classification to predict eGFR over a subject's adult life as a function of age and age-squared, which indicated ADPKD patients with different Mayo classification had different prognosis trajectories of eGFR decline. In this study, we refitted the mixed models and applied them in predicting a specific ADPKD participant's eGFR value based on a single observation of the age.

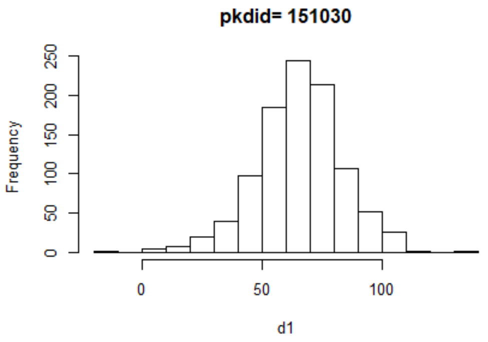
The three Chi-square test results led to several conclusions. The result of Chi-square test for observed outcomes showed that the observed outcomes of CKD status were different between the LTF and non-LTF participants. Because the LTF participants all had their last observed eGFR before Year 12, we assumed that some of them may enter worse CKD stage as they age (or reach Year 15), so that the outcome of the CKD status of LTF group may become more similar to the non-LTF group. Therefore, we conducted another chi-squared test, including participants with transplant or dialysis in stage 5 and included in the non-LTF group. This test indicated that the predicted outcome of CKD status of LTF participants were different from the observed outcome of CKD status of non-LTF participants. However, our data also showed that some of the non-LTF participants had kidney transplants or dialysis with their last observed eGFR values larger than 15 mm/min. In other words, some of the participants may receive transplant or dialysis at an earlier time (in CKD stage 1-4) rather than after they entered CKD stage 5. Based on this scenario, it is reasonable to assume that such a situation may also occur among LTF participants. In this case, we did the third chi-squared test between the two groups. In order to eliminate the influence of this factor, we excluded the participants who had transplant and dialysis from the non-LTF group. The test result showed there was no significant evidence that the predicted outcome of CKD status of LTF participants were different from the observed outcome of CKD status of non-LTF participants. Therefore, we concluded that in this situation, the outcome of the CKD status at Year 15 of the LTF participants was consistent with the outcome of non-LTF participants, so that the LTF group was considered to be the random subset of the cohort in CRISP.

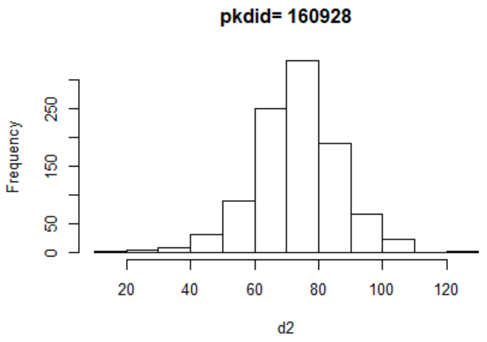
The bootstrap confidence intervals provided estimates of the variability of the predicted results. For the 44 LTF participants, only six (13.6%) of them yielded confidence intervals containing eGFR = 15 mm/min. As shown in Figure 2, 3, and 4, only six participants who had confidence interval containing eGFR = 15 mm/min may had inconsistent predicted CKD status in the 1000 bootstrap samples. Therefore, this approach seemed to yield sufficiently precise estimates of which subjects would reach stage 5 to then compare to those who were not lost to follow-up. Overall, results seem relatively consistent with the hypothesis that LTF participants represent a random subset of the total data set, i.e. LTF participants were lost from the cohort because of mostly random events, rather than non-random events.

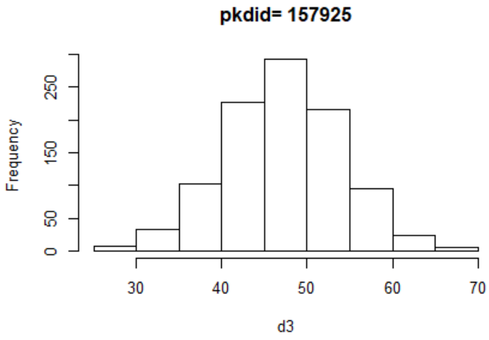
There were limitations to this study. First, the CRISP sample size had only a moderately large sample size. Other studies, including analysis of completed randomized trials, are ongoing at other sites with larger sample sizes. The data on eGFR was also irregularly spaced over different time periods, thus limiting precision of estimates at Year 15. There was also some error in defining who had completed the study, as the cut-off of Year 12 was relatively arbitrary. However, despite these limitations, the CRISP cohort provides a rather unique characterization of clinical characteristics and imaging measurements.

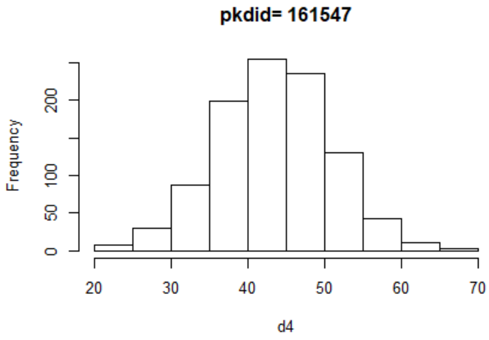
It is meaningful for public health to understand the patterns of disease developments, and the same is true for the ADPKD patients. Because lack of follow-up will affect the effectiveness of the study, it is important to obtain the information of the LTF participants as much as possible. Early understanding of the variability in patients with different disease risks could provide specific information for the development of disease, which is of great significance for proper prevention and treatment in the future.

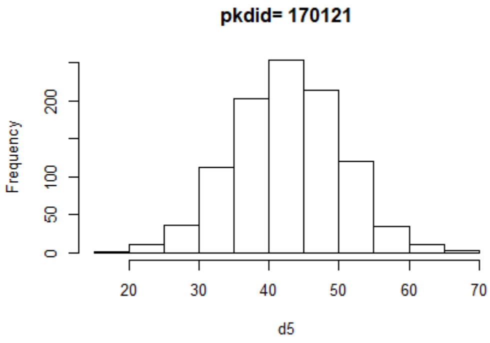
Appendix A Histograms of Bootstrap Estimated Results for Other LTF Participants

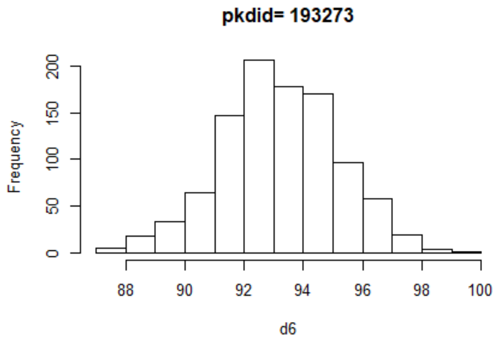

pkdid= 151030

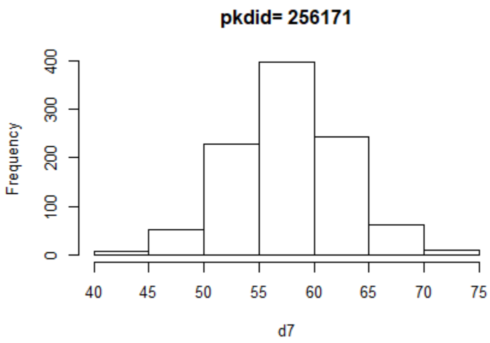

pkdid= 160928

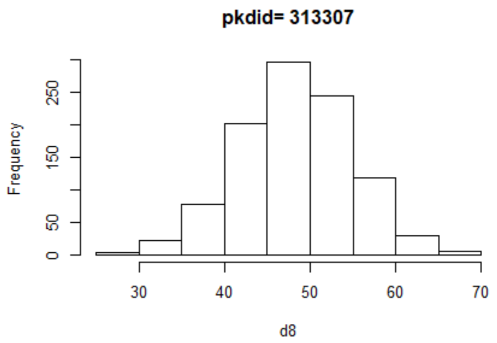

pkdid= 157925


pkdid= 161547

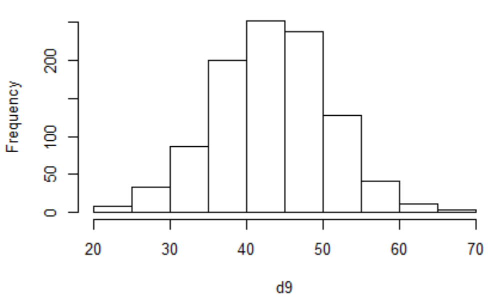

pkdid= 170121


pkdid= 193273

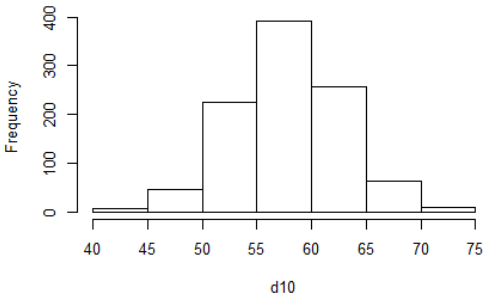

pkdid= 256171


pkdid= 313307

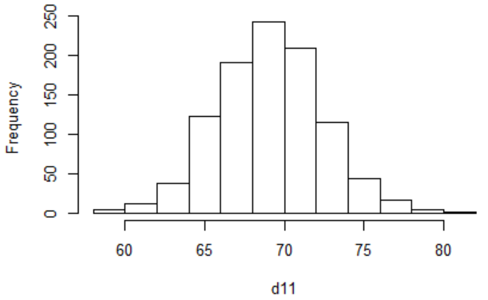
pkdid= 407648



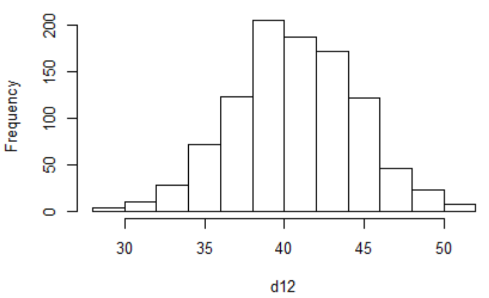
pkdid= 407841



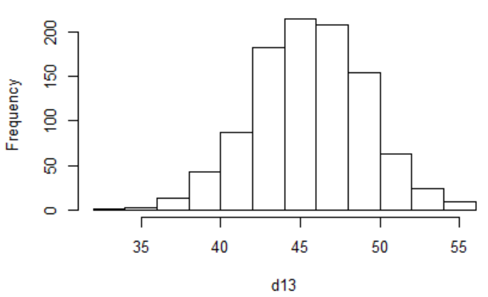
pkdid= 476972



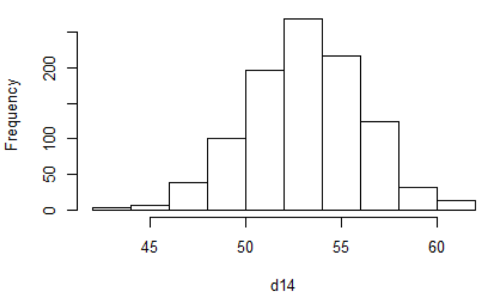
pkdid= 101585



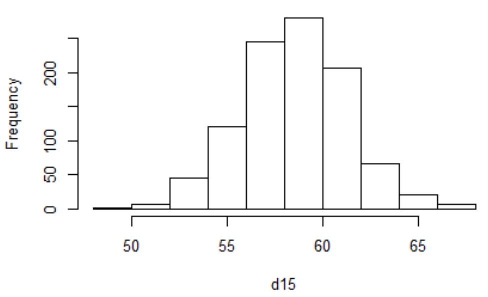
pkdid= 110080



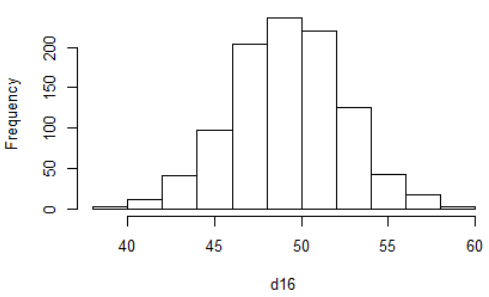
pkdid= 126133

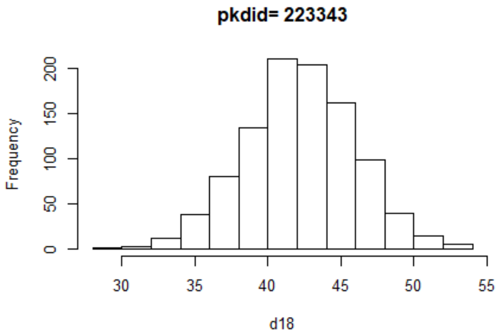


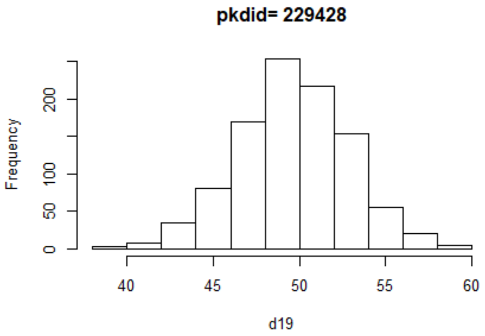
pkdid= 139486

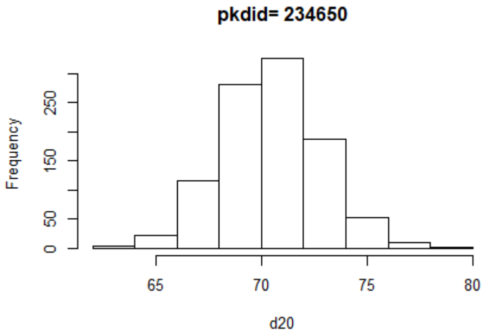


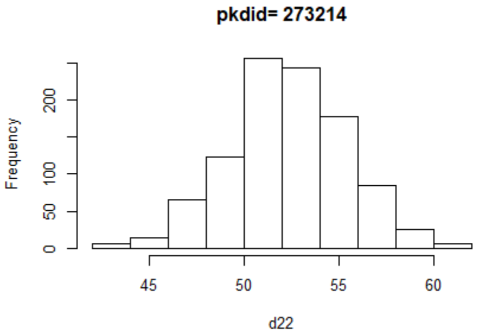
pkdid= 194105

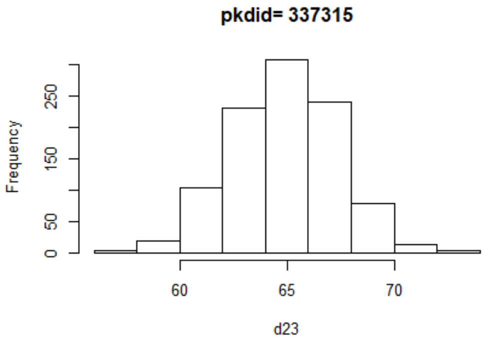


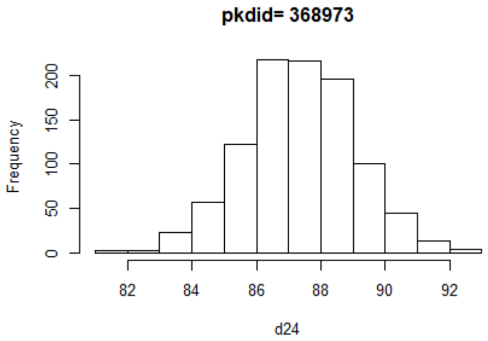

pkdid= 223343
Frequency
0
50
100
150
200
30
35
40
45
50
55
d18

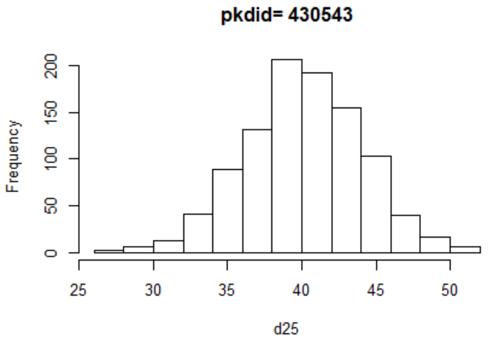

pkdid= 229428
Frequency
0
50
100
200
40
45
50
55
60
d19

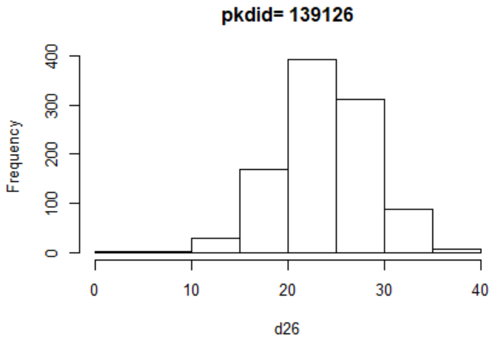

pkdid= 234650
Frequency
0
50
150
250
65
70
75
80
d20

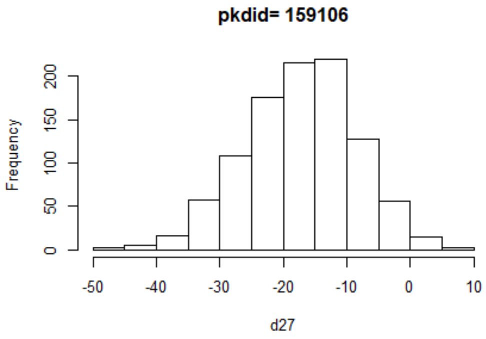

pkdid= 273214
Frequency
0
50
100
200
45
50
55
60
d22

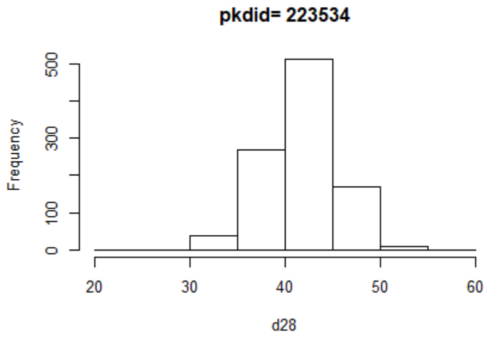

pkdid= 337315
Frequency
0
50
150
250
60
65
70
d23

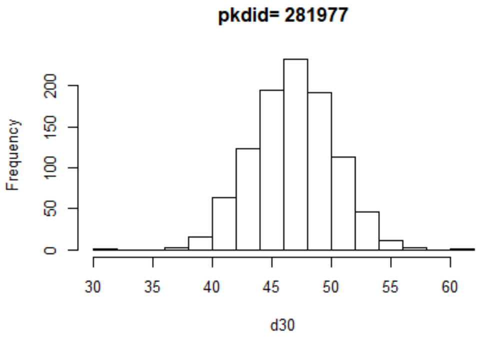

pkdid= 368973
Frequency
0
50
100
150
200
82
84
86
88
90
92
d24

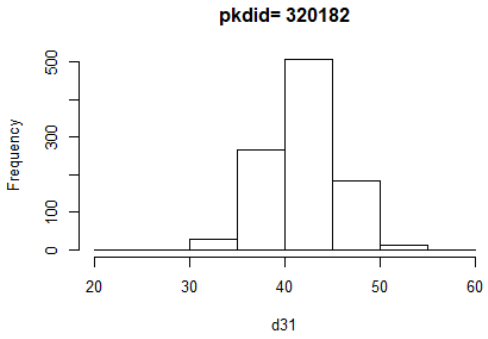

pkdid= 430543
Frequency
0
50
100
150
200
25
30
35
40
45
50
d25

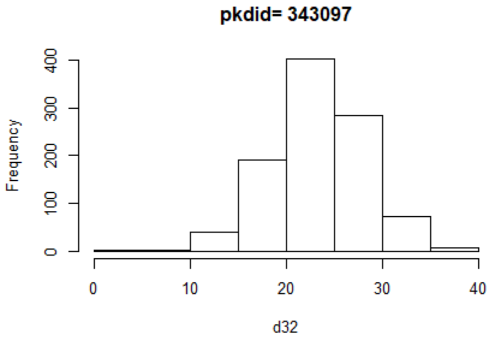

pkdid= 139126
Frequency
0
100
200
300
400
0
10
20
30
40
d26

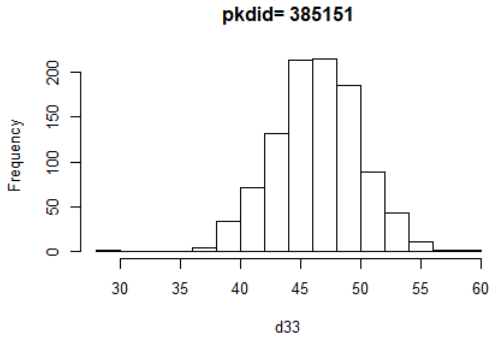

pkdid= 159106
Frequency
0
50
100
150
200
-50
-40
-30
-20
-10
0
10
d27

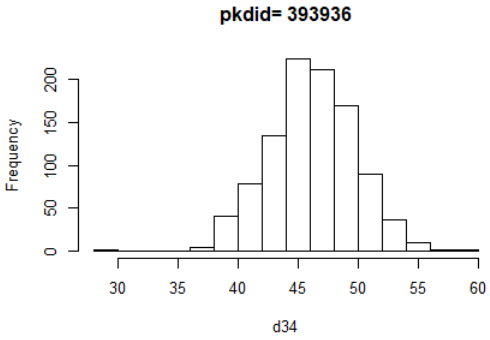

pkdid= 223534
Frequency
0
100
300
500
20
30
40
50
60
d28

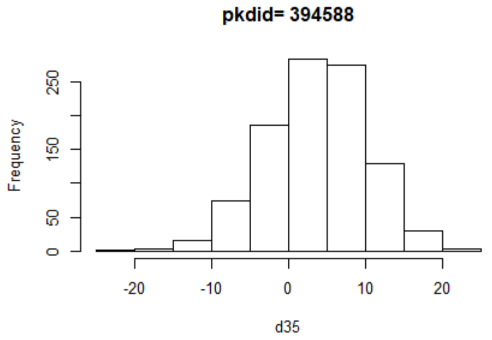

pkdid= 281977
Frequency
0
50
100
150
200
30
35
40
45
50
55
60
d30

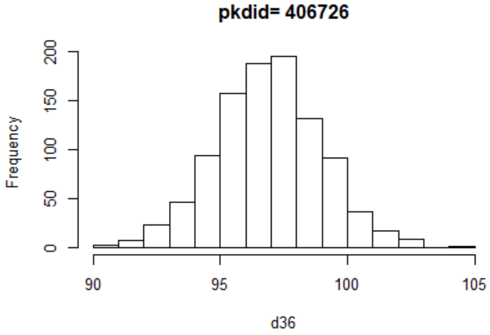

pkdid= 320182
Frequency
0
100
300
500
20
30
40
50
60
d31

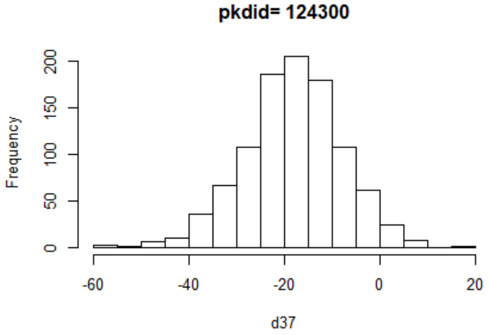

pkdid= 343097
Frequency
0
100
200
300
400
0
10
20
30
40
d32

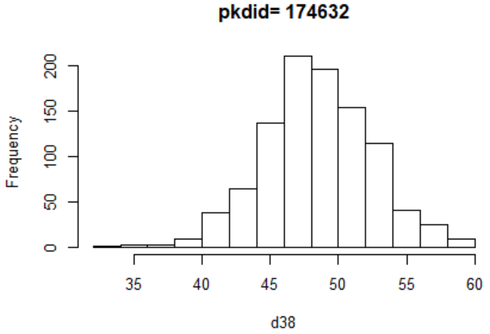

pkdid= 385151
Frequency
0
50
100
150
200
30
35
40
45
50
55
60
d33

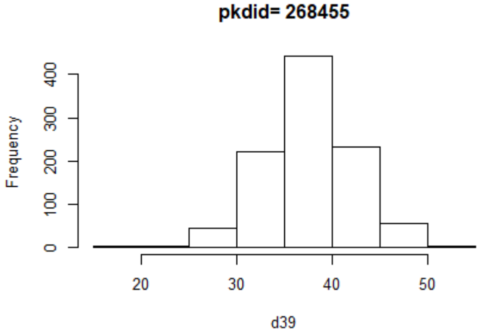

pkdid= 393936
Frequency
0
50
100
150
200
30
35
40
45
50
55
60
d34

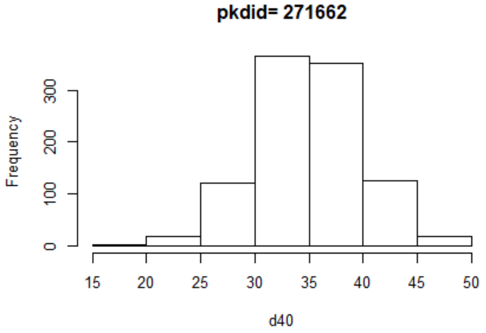

pkdid= 394588
Frequency
0
50
150
250
-20
-10
0
10
20
d35

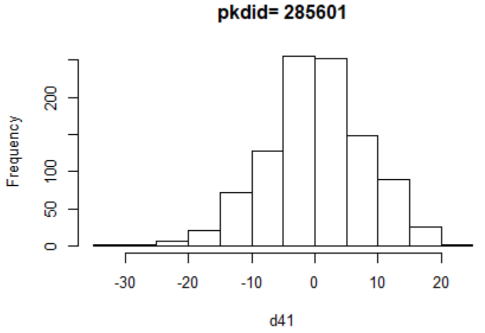

pkdid= 406726

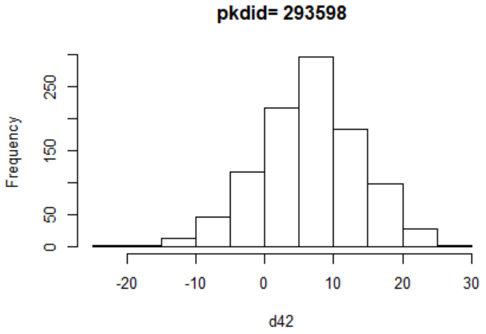

pkdid= 124300

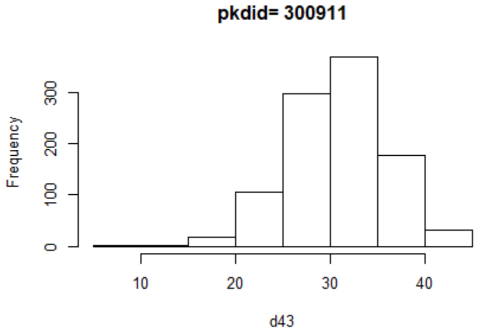

pkdid= 174632

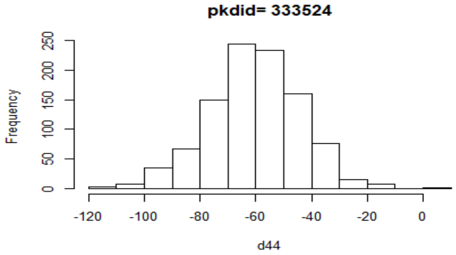

pkdid= 268455


pkdid= 271662


pkdid= 285601


pkdid= 293598


pkdid= 300911



Appendix Figure 1 Histograms of Bootstrap estimated results for other LTF participants

Appendix B Code Used in R

```
# Packages
library(car)
library(MASS)
library(lme4)
library(lmerTest)
library(MuMIn)
library(tidyr)
library(readxl)
library(ggplot2)

#Data sets
#"last egfr and IC from CRISP 3_with transplant and dialysis" is the data set contains the baseline info
#"data_b is a sorted data set including the indicator of the first observation and the last observation of a participant

data<-as.data.frame(read_excel("D:/Dr. Landsittel/CRISP/last egfr and IC from CRISP 3_with transplant and dialysis.xls"))
data0<-as.data.frame(read_excel("D:/Dr. Landsittel/CRISP/data_b.xls"))
data.ori<-as.data.frame(read_excel("D:/Dr. Landsittel/CRISP/CRISP I-III variables for Anni thesis.xls"))

data.first<-subset(data0,n1==1)
data.last<-subset(data0,n1==n2)

data.final<-merge(data.last, data, by="pkdid")
a<-data.final$ckd_epi
```

```
a<-as.numeric(a)


a<-na.omit(a)
mean(a)
sd(a)


data.o<-subset(data.ori,data.ori$vis==0)
data.final.2<-merge(data.final, data.o, by="p
kdid")


#Baseline and follow_up infomation


mean(data.final.2$httkv)
sd(data.final.2$httkv)


mean(data.final.2$ckd_epi.y)
sd(data.final.2$ckd_epi.y)


table(data.final$IC)
table(data.final$class)
table(data.final$transplant)
table(data.final$dialysis)


attach(data.final)
data.final$status[visc < 12 & ckd_epi >= 15 & transplant == 0 & dialysis == 0] <- "lost to follow up"
detach(data.final)


data.lf<-subset(data.final, data.final$status == "lost to follow up")
```

```
length(data.lf$pkdid)

attach(data.final)
data.final$status2[ckd_epi >=90 & transplant == 0 & dialysis == 0] <- "CKD stage 1" #
data.final$status2[visc < 12 & ckd_epi >=90 & transplant == 0 & dialysis == 0] <- "obs lost 1" #
data.final$status2[ckd_epi >=60 & ckd_epi < 90 & transplant == 0 & dialysis == 0] <- "CKD stage 2"  #
data.final$status2[visc < 12 & ckd_epi >=60 & ckd_epi < 90 & transplant == 0 & dialysis == 0] <- "obs lost 2"
data.final$status2[ckd_epi >=45 & ckd_epi < 60 & transplant == 0 & dialysis == 0] <- "CKD stage 3a" #
data.final$status2[visc < 12 & ckd_epi >=45 & ckd_epi < 60 & transplant == 0 & dialysis == 0] <- "obs lost 3a" #
data.final$status2[ckd_epi >=30 & ckd_epi < 45 & transplant == 0 & dialysis == 0] <- "CKD stage 3b" #
data.final$status2[visc < 12 & ckd_epi >=30 & ckd_epi < 45 & transplant == 0 & dialysis == 0] <- "obs lost 3b" #
data.final$status2[ckd_epi >=15 & ckd_epi < 30 & transplant == 0 & dialysis == 0] <- "CKD stage 4"  #
data.final$status2[visc < 12 & ckd_epi >=15 & ckd_epi < 30 & transplant == 0 & dialysis == 0] <- "obs lost 4"  #
data.final$status2[ckd_epi < 15] <- "CKD stage 5"
data.final$status2[transplant == 1] <- "Transplant"
data.final$status2[dialysis == 1] <- "Dialysis"
detach(data.final)

as.data.frame(table(data.final$status2))

#follow up > 12
nonltfdata<-subset(data.final,visc >=12)
attach(nonltfdata)
nonltfdata$status[ckd_epi >=90] <- "CKD stage 1"                    #32
```

```
nonltfdata$status[ckd_epi >=60 & ckd_epi < 90] <- "CKD stage 2"  #48
nonltfdata$status[ckd_epi >=45 & ckd_epi < 60] <- "CKD stage 3a" #17
nonltfdata$status[ckd_epi >=30 & ckd_epi < 45] <- "CKD stage 3b" #17
nonltfdata$status[ckd_epi >=15 & ckd_epi < 30] <- "CKD stage 4"  #19
nonltfdata$status[ckd_epi < 15] <- "CKD stage 5"                 #15
nonltfdata$status[transplant == 1] <- "transplant"
nonltfdata$status[dialysis == 1] <- "dialysis"
detach(nonltfdata)
as.data.frame(table(nonltfdata$status))

length(nonltfdata$pkdid)

daa<-subset(data.final,visc < 12)
length(daa$pkdid)
as.data.frame(table(daa$status2))



#lost to follow up
attach(data.lf)
data.lf$status[ckd_epi >=90] <- "CKD stage 1"                 #24
data.lf$status[ckd_epi >=60 & ckd_epi < 90] <- "CKD stage 2"  #19
data.lf$status[ckd_epi >=45 & ckd_epi < 60] <- "CKD stage 3a" #10
data.lf$status[ckd_epi >=30 & ckd_epi < 45] <- "CKD stage 3b" #1
data.lf$status[ckd_epi >=15 & ckd_epi < 30] <- "CKD stage 4"  #6
data.lf$status[ckd_epi < 15] <- "CKD stage 5"                 #29
data.lf$status[transplant == 1] <- "transplant"
data.lf$status[dialysis == 1] <- "dialysis"
detach(data.lf)
```

```
as.data.frame(table(data.lf$status))

length(data.lf$pkdid)

lfckd<-data.lf$ckd_epi
lfckd<-as.numeric(nlfckd)
lfckd<-na.omit(nlfckd)
mean(lfckd)
sd(lfckd)

data.nlf<-as.data.frame(read_excel("D:/Dr. Landsittel/CRISP/nltf.xls"))
data.nlf
nlfckd<-data.nlf$ckd_epi
nlfckd<-as.numeric(nlfckd)
nlfckd<-na.omit(nlfckd)
mean(nlfckd)
sd(nlfckd)



#Chi-square Test
#1.not LTF (combine T+D with stage 5) vs last observed status in LTF
chisq_data_1 <- data.frame(ckd_1_4 = c(112,80),
                           ckd_1_5 = c(44,0))

chisq.test(chisq_data_1)

#2.not LTF (combine T+D with stage 5) vs predicted outcome in LTF
chisq_data_2 <- data.frame(ckd_1_4 = c(112,80),
                    ckd_1_5 = c(37,7))


chisq.test(chisq_data_2)


#3.not LTF (do not combine T+D with stage 5) vs predicted outcome in LTF
chisq_data_3 <- data.frame(ckd_1_4 = c(112,41),
                    ckd_1_5 = c(37,7))


chisq.test(chisq_data_3)


#Mixed models
load("D:/Dr. Landsittel/CRISP/Data/data_Anni.RData")


# polynomial model for class A
CKD_epi1<-mydata1[mydata1$class==1,]$ckd_epi
age1<-mydata1[mydata1$class==1,]$age
pkdid1<-mydata1[mydata1$class==1,]$pkdid


age1_2=age1^2
CKD_epi1_p<- lmer(CKD_epi1 ~age1+age1_2+( 1|pkdid1))
summary(CKD_epi1_p)


# polynomial model for class B
CKD_epi2<-mydata1[mydata1$class==2,]$ckd_epi
age2<-mydata1[mydata1$class==2,]$age
pkdid2<-mydata1[mydata1$class==2,]$pkdid


age2_2=age2^2
```

```
CKD_epi2_p<- lmer(CKD_epi2 ~age2+age2_2+( 1|pkdid2))
summary(CKD_epi2_p)

# polynomial model for class C
CKD_epi3<-mydata1[mydata1$class==3,]$ckd_epi
age3<-mydata1[mydata1$class==3,]$age
pkdid3<-mydata1[mydata1$class==3,]$pkdid

age3_2<-age3^2
CKD_epi3_p<- lmer(CKD_epi3 ~age3+age3_2+( 1|pkdid3))
summary(CKD_epi3_p)

# polynomial model for class D
CKD_epi4<-mydata1[mydata1$class==4,]$ckd_epi
age4<-mydata1[mydata1$class==4,]$age
pkdid4<-mydata1[mydata1$class==4,]$pkdid

age4_2<-age4^2
CKD_epi4_p<- lmer(CKD_epi4 ~age4+age4_2+( 1|pkdid4))
summary(CKD_epi4_p)

# polynomial model for class E
CKD_epi5<-mydata1[mydata1$class==5,]$ckd_epi
age5<-mydata1[mydata1$class==5,]$age
pkdid5<-mydata1[mydata1$class==5,]$pkdid

age5_2<-age5^2
CKD_epi5_p<- lmer(CKD_epi5 ~age5+age5_2+( 1|pkdid5))
```

```
summary(CKD_epi5_p)



#coefficient for individuals
ICa_coef<-coef(CKD_epi1_p)$pkdid
ICb_coef<-coef(CKD_epi2_p)$pkdid
ICc_coef<-coef(CKD_epi3_p)$pkdid
ICd_coef<-coef(CKD_epi4_p)$pkdid
ICe_coef<-coef(CKD_epi5_p)$pkdid

data<-as.data.frame(read_excel("D:/Dr. Landsittel/CRISP/last egfr and IC from CRISP 3_with transplant and dialysis.xls"))
data0<-as.data.frame(read_excel("D:/Dr. Landsittel/CRISP/data_b.xls"))

data.first<-subset(data0,n1==1)
data.last<-subset(data0,n1==n2)
data.final<-merge(data.last, data, by="pkdid")

attach(data.final)
data.final$status[visc < 12 & ckd_epi >= 15 & transplant == 0 & dialysis == 0] <- "lost to follow up"
detach(data.final)

data.lf<-subset(data.final, data.final$status == "lost to follow up")
length(data.lf$pkdid)

data.a<-merge(data.lf, data.f.final, by="pkdid")
data.a$ckd_epi<-as.numeric(data.a$ckd_epi)
data.f.final<-merge(data.first, data.lf, by="pkdid")

#p$ckd_epi<-as.numeric(levels(p$ckd_epi)[p$ckd_epi])

data.f.final$ckd_epi.x<-as.numeric(data.f.final$ckd_epi.x)



et<-data.f.final$baseline_age+15

et2<-et^2


data.lf<-cbind(data.lf,et)

data.f.final<-merge(data.first, data.lf, by="pkdid")


#"Ic_coef" is the sorted coefficients of the participants

ic_coef<-as.data.frame(read_excel("D:/Dr. Landsittel/CRISP/coef.xls"))

lf_coef<-subset(ic_coef,pkdid %in% data.a$pkdid)



#"pp" is used to calculate the predicted values

pp<-data.frame(

  pkdid=data.a$pkdid,

  baseline_age=data.a$baseline_age.x,

  ckd_epi=as.numeric(data.a$ckd_epi),

  IC=data.a$IC.y,

  et=et,

  et2=et2

)

pp<-merge(pp,lf_coef,by="pkdid")
```

```
predict<-pp$intercept + pp$age*pp$et + pp$age2*pp$et2


#"p" is the data set contains the baseline info and predicted info
p<-data.frame(
  pkdid=pp$pkdid,
  baseline_age=pp$baseline_age,
  ckd_epi=as.numeric(pp$ckd_epi),
  intercept=pp$intercept,
  age=pp$age,
  age2=pp$age2,
  IC=pp$IC.x,
  et=pp$et,
  et2=pp$et2,
  predict_result=predict
)


#Bootstrap ######################################################


#p<-as.data.frame(read_excel("D:/Dr. Landsittel/CRISP/p.xls"))
#p$ckd_epi<-as.numeric(levels(p$ckd_epi)[p$ckd_epi])
#p$ckd_epi<-as.numeric(p$ckd_epi)
#str(p)



data0<-as.data.frame(read_excel("D:/Dr. Landsittel/CRISP/data_b.xls"))
age2<-(data0$age)^2
data0<-cbind(data0,age2)
```

```
data.bs<-cbind(data0$pkdid, data0$ckd_epi, data0$age, data0$age2, data0$class)

data.bs<-data.frame(
  pkdid=data0$pkdid,
  ckd_epi=data0$ckd_epi,
  age=data0$age,
  age2=data0$age2,
  class=data0$class)
data.bs$ckd_epi<-as.numeric(levels(data.bs$ckd_epi)[data.bs$ckd_epi])
str(data.bs)

table(data.bs$class)

LTF_data<-data.frame(
  pkdid=p$pkdid,
  et=p$et,
  et2=p$et2,
  IC=p$IC
)

s<-matrix()

for(i in 1:236){
  s[i]<-subset(data.bs,pkdid==data.final[i,1])
}
s[1]

for(i in 1:236){
  s[[i]]<-subset(data.bs,pkdid==data.final[i,1])
}
s[[1]]

#Start Bootstrapping
nboot <-1000  #number of bootstrap samples
m<-matrix(0, 44, 1000)
bootstrap.m<-matrix()
unl.bs<-matrix()

set.seed(04202020)

for(i in 1:nboot){
  bootstrap.m[i]<-sample(s,size=236,replace=TRUE)
}

for(i in 1:nboot){
  bootstrap.m[[i]]<-sample(s,size=236,replace=TRUE)
  unl.bs[i]<-unlist(bootstrap.m[i])
}
  bootstrap1 <- matrix()
  bootstrap1 <- as.data.frame(bootstrap.m[[1]][1])
  for (i in 2:236) {
    x <- as.data.frame(bootstrap.m[[1]][i])
    bootstrap1 <- rbind(bootstrap1, x)
  }
  #bs contains all the 1000 BS samples
  bs <- list()
  for(i in 1:nboot){
    boots <- as.data.frame((bootstrap.m[[i]][1]))
    for(j in 2:236){
      x <- as.data.frame(bootstrap.m[[i]][j])
      boots <- rbind(boots, x)
    }
    bs[[i]] <- boots
  }



#To obtain the coefficients of the 5 mixed model refitted by the 1000 samples
s1<-matrix()
s2<-matrix()
s3<-matrix()
s4<-matrix()
s5<-matrix()

for(i in 1:nboot){
s1[i]<-subset(bs[[i]],bs[[i]]$class==1)
s2[i]<-subset(bs[[i]],bs[[i]]$class==2)
s3[i]<-subset(bs[[i]],bs[[i]]$class==3)
s4[i]<-subset(bs[[i]],bs[[i]]$class==4)
s5[i]<-subset(bs[[i]],bs[[i]]$class==5)
  }
```

```
for(i in 1:nboot){
  s1[[i]]<-subset(bs[[i]],bs[[i]]$class==1)
  s2[[i]]<-subset(bs[[i]],bs[[i]]$class==2)
  s3[[i]]<-subset(bs[[i]],bs[[i]]$class==3)
  s4[[i]]<-subset(bs[[i]],bs[[i]]$class==4)
  s5[[i]]<-subset(bs[[i]],bs[[i]]$class==5)
}



#coef1-coef5 contains the coefficents of 1000 mixed models for each Mayo classification

coef1<-data.frame()
for(i in 1:nboot){
coef1<- rbind(coef1,coef(summary(lmer(ckd_epi ~age+age2+( 1|pkdid), data=s1[[i]])))[,1])
}

coef2<-data.frame()
for(i in 1:nboot){
  coef2<- rbind(coef2,coef(summary(lmer(ckd_epi ~age+age2+( 1|pkdid), data=s2[[i]])))[,1])
}

coef3<-data.frame()
for(i in 1:nboot){
  coef3<- rbind(coef3,coef(summary(lmer(ckd_epi ~age+age2+( 1|pkdid), data=s3[[i]])))[,1])
}

coef4<-data.frame()
```

```
for(i in 1:nboot){
  coef4<- rbind(coef4,coef(summary(lmer(ckd_epi ~age+age2+( 1|pkdid), data=s4[[i]])))[,1])
}

coef5<-data.frame()
for(i in 1:nboot){
  coef5<- rbind(coef5,coef(summary(lmer(ckd_epi ~age+age2+( 1|pkdid), data=s5[[i]])))[,1])
}

m1<-subset(LTF_data,LTF_data$IC=="A")
m2<-subset(LTF_data,LTF_data$IC=="B")
m3<-subset(LTF_data,LTF_data$IC=="C")
m4<-subset(LTF_data,LTF_data$IC=="D")
m5<-subset(LTF_data,LTF_data$IC=="E")


#To obtain the BS estimated eGFR values for 44 LTF participants

pre1<-c()
d1<-matrix()
for(i in 1:nboot){
  pre1[i]<-coef1[i,1]+ m1[1,2]*coef1[i,2] + m1[1,3]*coef1[i,3]
  d1<-cbind(d1,pre1[i])
}
d1<-as.vector(d1)
d1<-sort(d1)

pre2<-c()
```

```
d2<-matrix()
for(i in 1:nboot){
  pre2[i]<-coef1[i,1]+ m1[2,2]*coef1[i,2] + m1[2,3]*coef1[i,3]
  d2<-cbind(d2,pre2[i])
}
d2<-as.vector(d2)
d2<-sort(d2)

pre3<-c()
d3<-matrix()
for(i in 1:nboot){
  pre3[i]<-coef2[i,1]+ m2[1,2]*coef2[i,2] + m2[1,3]*coef2[i,3]
  d3<-cbind(d3,pre3[i])
}
d3<-as.vector(d3)
d3<-sort(d3)

……

pre44<-c()
d44<-matrix()
for(i in 1:nboot){
  pre44[i]<-coef5[i,1]+ m5[8,2]*coef5[i,2] + m5[8,3]*coef5[i,3]
  d44<-cbind(d44,pre44[i])
}
d44<-as.vector(d44)
d44<-sort(d44)
```

```
#To obtain the 95% confidence interval


ci1<-c(d1[25],d1[975])

ci2<-c(d2[25],d2[975])

ci3<-c(d3[25],d3[975])

……

ci44<-c(d44[25],d44[975])




#95% confidence interval

boot.ci<-rbind(ci1,ci2,ci3,ci4,ci5,ci6,ci7,ci8,ci9,ci10,

          ci11,ci12,ci13,ci14,ci15,ci16,ci17,ci18,ci19,ci20,

          ci21,ci22,ci23,ci24,ci25,ci26,ci27,ci28,ci29,ci30,

          ci31,ci32,ci33,ci34,ci35,ci36,ci37,ci38,ci39,ci40,

          ci41,ci42,ci43,ci44)


p<-p[

  order( p$IC ),

  ]

p


d<-rbind(d1,d2,d3,d4,d5,d6,d7,d8,d9,d10,

      d11,d12,d13,d14,d15,d16,d17,d18,d19,d20,

      d21,d22,d23,d24,d25,d26,d27,d28,d29,d30,

      d31,d32,d33,d34,d35,d36,d37,d38,d39,d40,

      d41,d42,d43,d44)
```

```
#To obtain the corresponding CKD stages
c.t <- cut(d, breaks = c(-150, 15, 30, 45, 60, 90, 150))
attr(c.t , 'levels')
attr(c.t , 'class')
ckd_stage<-ordered(c.t , labels = c('CKD 1', 'CKD 2', 'CKD 3a', 'CKD 3b', 'CKD 4', 'CKD 5'))
ckd_stage<-(na.omit(ckd_stage))
ckd_stage
#write.csv(ckd_stage,"D:/Dr. Landsittel/CRISP/ckd_stage.csv",row.names = FALSE)


c1<-ckd_stage[1:1000]
c2<-ckd_stage[1001:2000]
c3<-ckd_stage[2001:3000]
……
c44<-ckd_stage[43001:44000]


t1<-table(c1)
t2<-table(c2)
t3<-table(c3)
……
t44<-table(c44)


ckd_dis<-as.data.frame(rbind(t1,t2,t3,t4,t5,t6,t7,t8,t9,t10,
                             t11,t12,t13,t14,t15,t16,t17,t18,t19,t20,
              t21,t22,t23,t24,t25,t26,t27,t28,t29,t30,
              t31,t32,t33,t34,t35,t36,t37,t38,t39,t40,
              t41,t42,t43,t44))

#To obtain the % of CKD stage for 1000 eGFR values
ckd_per<-ckd_dis/1000*100
ckd_per

p<-p[
  order( p$IC ),
  ]
p

p<-cbind(p,ckd_per)
p
str(p)



d<-cbind(pkdid=p$pkdid, IC=p$IC, d)
d<-as.data.frame(d)
d



p<-p[
  order( p$pkdid ),
  ]
p
```

```
p<-p[
  order( p$IC ),
  ]
p

#Histograms of predicted eGFR values

#Examples
hist(d29,xlab="Predicted eGFR values",main=paste("pkdid=",p$pkdid[29]))#236202(1)
hist(d17,xlab="Predicted eGFR values",main=paste("pkdid=",p$pkdid[17]))#195310(2)
hist(d21,xlab="Predicted eGFR values",main=paste("pkdid=",p$pkdid[21]))#244111(3)


#Appendix
par(mfrow=c(2,2))
hist(d1,main=paste("pkdid=",p$pkdid[1]))
hist(d2,main=paste("pkdid=",p$pkdid[2]))
hist(d3,main=paste("pkdid=",p$pkdid[3]))
hist(d4,main=paste("pkdid=",p$pkdid[4]))
par(mfrow=c(1,1))

par(mfrow=c(2,2))
hist(d5,main=paste("pkdid=",p$pkdid[5]))
hist(d6,main=paste("pkdid=",p$pkdid[6]))
hist(d7,main=paste("pkdid=",p$pkdid[7]))
hist(d8,main=paste("pkdid=",p$pkdid[8]))
par(mfrow=c(1,1))
par(mfrow=c(2,2))

hist(d9,main=paste("pkdid=",p$pkdid[9]))

hist(d10,main=paste("pkdid=",p$pkdid[10]))

hist(d11,main=paste("pkdid=",p$pkdid[11]))

hist(d12,main=paste("pkdid=",p$pkdid[12]))

par(mfrow=c(1,1))


par(mfrow=c(2,2))

hist(d13,main=paste("pkdid=",p$pkdid[13]))

hist(d14,main=paste("pkdid=",p$pkdid[14]))

hist(d15,main=paste("pkdid=",p$pkdid[15]))

hist(d16,main=paste("pkdid=",p$pkdid[16]))

par(mfrow=c(1,1))



par(mfrow=c(2,2))

hist(d18,main=paste("pkdid=",p$pkdid[18]))

hist(d19,main=paste("pkdid=",p$pkdid[19]))

hist(d20,main=paste("pkdid=",p$pkdid[20]))

hist(d22,main=paste("pkdid=",p$pkdid[22]))

par(mfrow=c(1,1))



par(mfrow=c(2,2))

hist(d23,main=paste("pkdid=",p$pkdid[23]))

hist(d24,main=paste("pkdid=",p$pkdid[24]))

hist(d25,main=paste("pkdid=",p$pkdid[25]))

hist(d26,main=paste("pkdid=",p$pkdid[26]))
```

```
par(mfrow=c(1,1))


par(mfrow=c(2,2))
hist(d27,main=paste("pkdid=",p$pkdid[27]))
hist(d28,main=paste("pkdid=",p$pkdid[28]))
hist(d30,main=paste("pkdid=",p$pkdid[30]))
hist(d31,main=paste("pkdid=",p$pkdid[31]))
par(mfrow=c(1,1))


par(mfrow=c(2,2))
hist(d32,main=paste("pkdid=",p$pkdid[32]))
hist(d33,main=paste("pkdid=",p$pkdid[33]))
hist(d34,main=paste("pkdid=",p$pkdid[34]))
hist(d35,main=paste("pkdid=",p$pkdid[35]))
par(mfrow=c(1,1))


par(mfrow=c(2,2))
hist(d36,main=paste("pkdid=",p$pkdid[36]))
hist(d37,main=paste("pkdid=",p$pkdid[37]))
hist(d38,main=paste("pkdid=",p$pkdid[38]))
hist(d39,main=paste("pkdid=",p$pkdid[39]))
par(mfrow=c(1,1))


par(mfrow=c(2,2))
hist(d40,main=paste("pkdid=",p$pkdid[40]))
hist(d41,main=paste("pkdid=",p$pkdid[41]))
hist(d42,main=paste("pkdid=",p$pkdid[42]))
hist(d43,main=paste("pkdid=",p$pkdid[43]))
```

```
par(mfrow=c(1,1))


par(mfrow=c(2,2))
hist(d44,main=paste("pkdid=",p$pkdid[44]))
par(mfrow=c(1,1))



#To obtain the average values of the 1000 BS estimated results

bm1<-mean(d1)
bm2<-mean(d2)
bm3<-mean(d3)
……
bm44<-mean(d44)

bs_mean<-c(bm1,bm2,bm3,bm4,bm5,bm6,bm7,bm8,bm9,bm10,
           bm11,bm12,bm13,bm14,bm15,bm16,bm17,bm18,bm19,bm20,
           bm21,bm22,bm23,bm24,bm25,bm26,bm27,bm28,bm29,bm30,
           bm31,bm32,bm33,bm34,bm35,bm36,bm37,bm38,bm39,bm40,
           bm41,bm42,bm43,bm44)
bs_mean



p<-p[
  order( p$IC ),
  ]
p


data.results<-p
data.results<-cbind(data.results,bs_mean,boot.ci)

attach(data.results)
data.results$bs_mean_status[bs_mean >=90] <- "CKD stage 1"                   #24
data.results$bs_mean_status[bs_mean >=60 & bs_mean < 90] <- "CKD stage 2"  #19
data.results$bs_mean_status[bs_mean >=45 & bs_mean < 60] <- "CKD stage 3a" #10
data.results$bs_mean_status[bs_mean >=30 & bs_mean < 45] <- "CKD stage 3b" #1
data.results$bs_mean_status[bs_mean >=15 & bs_mean < 30] <- "CKD stage 4"  #6
data.results$bs_mean_status[bs_mean < 15] <- "CKD stage 5"                   #29
detach(data.results)
cbind(data.results$predicted_status, data.results$bs_mean_status)

############Data sorting part
#data.results is the "p" data set with some more variables

attach(data.results)
data.results$lastobs_status[ckd_epi >=90] <- "CKD stage 1"                   #24
data.results$lastobs_status[ckd_epi >=60 & ckd_epi < 90] <- "CKD stage 2"  #19
data.results$lastobs_status[ckd_epi >=45 & ckd_epi < 60] <- "CKD stage 3a" #10
data.results$lastobs_status[ckd_epi >=30 & ckd_epi < 45] <- "CKD stage 3b" #1
data.results$lastobs_status[ckd_epi >=15 & ckd_epi < 30] <- "CKD stage 4"  #6
data.results$lastobs_status[ckd_epi < 15] <- "CKD stage 5"                   #29
detach(data.results)
```

```
attach(data.results)

data.results$predicted_status[predict_result >=90] <- "CKD stage 1"                    #24

data.results$predicted_status[predict_result >=60 & predict_result < 90] <- "CKD stage 2"  #19

data.results$predicted_status[predict_result >=45 & predict_result < 60] <- "CKD stage 3a" #10

data.results$predicted_status[predict_result >=30 & predict_result < 45] <- "CKD stage 3b" #1

data.results$predicted_status[predict_result >=15 & predict_result < 30] <- "CKD stage 4"  #6

data.results$predicted_status[predict_result < 15] <- "CKD stage 5"                    #29

detach(data.results)


data.results<-data.results[

  order( data.results$IC ),

  ]


data.results<-cbind(data.results,bs_mean)


data.results<-data.results[

  order( data.results$pkdid ),

  ]

data.results
```

Bibliography

1. Torres, V.E., P.C. Harris, and Y. Pirson, *Autosomal dominant polycystic kidney disease.* The Lancet, 2007. **369**(9569): p. 1287-1301.
2. Harris, P.C., *Molecular basis of polycystic kidney disease: PKD1, PKD2 and PKHD1.* Current opinion in nephrology and hypertension, 2002. **11**(3): p. 309-314.
3. Harris, P.C. and V.E. Torres, *Polycystic kidney disease.* Annual review of medicine, 2009. **60**: p. 321-337.
4. Levey, A.S. and J. Coresh, *Chronic kidney disease.* The lancet, 2012. **379**(9811): p. 165-180.
5. Stats, F., *National chronic kidney disease fact sheet, 2017.* US Department of Health and Human Services, Centers for Disease Control and Prevention, 2017.
6. Wang, H., et al., *Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015.* The lancet, 2016. **388**(10053): p. 1459-1544.
7. Levey, A.S., et al., *K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification.* American Journal of Kidney Diseases, 2002. **39**(2 SUPPL. 1).
8. Valente, M.A., et al., *The Chronic Kidney Disease Epidemiology Collaboration equation outperforms the Modification of Diet in Renal Disease equation for estimating glomerular filtration rate in chronic systolic heart failure.* European journal of heart failure, 2014. **16**(1): p. 86-94.
9. Silveiro, S.P., et al., *Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation pronouncedly underestimates glomerular filtration rate in type 2 diabetes.* Diabetes care, 2011. **34**(11): p. 2353-2355.
10. Pietrangelo, A. *Stages of Chronic Kidney Disease*. 2020; Available from: https://www.healthline.com/health/ckd-stages.
11. Alan, S., et al., *Long-term trajectory of kidney function in autosomal-dominant polycystic kidney disease.* Kidney international, 2019. **95**(5): p. 1253-1261.
12. Chapman, A.B., et al., *Kidney volume and functional outcomes in autosomal dominant polycystic kidney disease.* Clinical Journal of the American Society of Nephrology, 2012. **7**(3): p. 479-486.
13. 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 international, 2003. **64**(3): p. 1035-1045.
14. (NIH), N.I.o.H. and N.I.o.D.a.D.a.K.D. (NIDDK). *Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP)*. 2001 [cited 2001; Available from: https://repository.niddk.nih.gov/studies/crisp/.
15. Mirman, D., *Growth curve analysis and visualization using R*. 2016: CRC press.
16. Grantham, J.J., *Autosomal dominant polycystic kidney disease.* New England Journal of Medicine, 2008. **359**(14): p. 1477-1485.
17. Irazabal, M.V., et al., *Imaging classification of autosomal dominant polycystic kidney disease: a simple model for selecting patients for clinical trials.* Journal of the American Society of Nephrology, 2015. **26**(1): p. 160-172.

18. Brownlee, J., *How to Calculate Bootstrap Confidence Intervals For Machine Learning Results in Python*. 2017.