PREDICTING POTENTIAL BIOMARKERS FOR EARLY DIABETIC NEPHROPATHY IN CHILDREN WITH TYPE I DIABETES MELLITUS

by

Haoyi Fu

BS, Nanjing Agricultural University, China, 2015

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Haoyi Fu

It was defended on

June 21st, 2017

and approved by

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

Ingrid M Libman, MD, PhD, Associate Professor
Department of Pediatrics, Children's Hospital of Pittsburgh
University of Pittsburgh Medical Center

Thesis Advisor: Vincent C Arena, PhD, Associate Professor
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Vincent C. Arena, PhD

PREDICTING POTENTIAL BIOMARKERS FOR EARLY DIABETIC NEPHROPATHY IN CHILDREN WITH TYPE I DIABETES MELLITUS

Haoyi Fu, MS

University of Pittsburgh, 2017

ABSTRACT

Type 1 Diabetes Mellitus (T1D) is a common form of Diabetes Mellitus worldwide and can cause long-term complications, especially in children. Diabetic nephropathy (DN) is the leading cause of mortality in T1D. The non-invasive gold standard for screening, monitoring, and predicting progression of DN is the assessment of albuminuria. However, it has been shown to lack sensitivity and specificity for early pathological manifestations of the disease. Other biomarkers including α-klotho, serum uric acid and estimated glomerular filtration rate (GFR) might potentially have a better ability to detect onset of DN earlier. The goal of this study is to gain a better understanding of how these biomarkers are associated with demographic and clinical characteristics in children.

Data from 97 children, age 10 or more years with a T1D duration of at least 2 years, were collected at Children's Hospital of Pittsburgh over a 4 month period. Correlations and univariable regression models were built to detect whether significant relationships between these biomarkers and demographic and clinical predictors were present. Multivariate regression models for each of the biomarkers were constructed and the cross-validation method was used to validate the models. After selecting the final models, linear regression assumptions were checked and model diagnostics were performed to detect problematic data points.

The final model for α-klotho contained the variables of hemoglobin A1c, growth velocity, triglycerides, total cholesterol, HDL and central obesity. For estimated GFR, the model included hemoglobin A1c, diastolic blood pressure percentile, growth velocity, albumin creatinine ratio (ACR), creatinine, total cholesterol and central obesity. The final model for serum uric acid included hemoglobin A1c, diabetic duration years, age, ACR, creatinine, triglycerides, total cholesterol, HDL, central obesity and waist percentile. Model fit criteria for all three multivariate models were largely improved compared to univariable models. Model diagnostics showed few problematic data points and linear regression assumptions for all three best models were not violated.

Public Health Significance: Although these biomarkers have been studied in adults with respect to screening, monitoring, and predicting progression of DN, less work has been done in pediatric populations. The work here provides a better understanding of the relationship between these biomarkers, and the demographic and clinical characteristics of children with T1D. The regression model validation techniques employed provide models that are not overly optimistic with respect to prediction. These methods are also more appropriate for studies with smaller sample sizes as found in this study.

TABLE OF CONTENTS

PREFACE XIV

1.0 INTRODUCTION 1

1.1 TYPE 1 DIABETES MELLITUS 1

1.2 DIABETIC NEPHROPATHY 2

1.3 ALPHA-KLOTHO 3

1.4 ESTIMATED GFR 3

1.5 SERUM URIC ACID 4

1.6 GOALS 4

2.0 METHODS 6

2.1 STUDY DESIGN AND SUBJECTS 6

2.2 INDEPENDENT VARIABLES 7

2.2.1 Demographic variables 7

2.2.2 Clinical variables 8

2.3 DATA MANAGEMENT 10

2.4 STATISTICAL ANALYSIS 10

2.4.1 Correlation 11

2.4.2 Univariable regression 12

2.4.3 Multiple regression 12

2.4.4 Model selection methods ... 12

2.4.5 Model selection criteria ... 14

2.4.6 Model validation techniques ... 16

2.4.7 Problematic points checking ... 19

2.4.8 Software ... 20

3.0 RESULTS ... 21

3.1 DESCRIPTIVE CHARACTERISITICS ... 21

3.1.1 Continuous variables ... 21

3.1.2 Categorical variables ... 22

3.2 CORRELATIONS ... 24

3.2.1 α-klotho and independent variables ... 24

3.2.2 Estimated GFR and independent variables ... 25

3.2.3 Serum uric acid and independent variables ... 26

3.3 TRANSFORMATIONS OF DEPENDENT VARIABLES ... 28

3.4 UNIVARIABLE REGRESSION ... 29

3.4.1 Log (α-klotho) ... 29

3.4.2 Log (eGFR) ... 30

3.4.3 Log (serum uric acid) ... 32

3.5 MULTIPLE REGRESSION ... 33

3.5.1 Variable selection criteria ... 33

3.5.2 Cross Validation ... 33

3.6 REGRESSION DIAGNOSTICS ... 44

3.6.1 Collinearity ... 44

3.6.2 Linear regression assumptions checking 46

3.6.3 Problematic points 51

3.7 ACR MODELS 56

3.7.1 Transformation of ACR 56

3.7.2 ACR model with potential biomarkers 57

3.7.3 ACR model with biomarkers and other variables 58

4.0 DISCUSSION AND CONCLUSION 60

APPENDIX A: SUPPLEMENTAL RESULTS.. 64

APPENDIX B: RELEVANT SAS CODES... 85

BIBLIOGRAPHY 96

LIST OF TABLES

Table 1. Descriptive statistics for continuous variables.. 22

Table 2. Descriptive statistics for categorical variables.. 23

Table 3. Correlations between α-klotho and other independent variables.................................... 25

Table 4. Correlations between eGFR and other independent variables....................................... 26

Table 5. Correlations between serum uric acid and other independent variables......................... 27

Table 6. Univariable regression models for log α-klotho 30

Table 7. Univariable regression models for log eGFR 31

Table 8. Univariable regression models for log serum uric acid 32

Table 9. Selected variables for different model selection criteria and k-fold cross-validation for log(α-klotho) model 35

Table 10. Comparisons of model selection criteria for log (α-klotho) model 35

Table 11. Model selection criteria with or without central obesity and waist% for α-klotho model 37

Table 12. Model selection criteria without central obesity and waist% for non-missing dataset for α-klotho model.. 37

Table 13. Parameter estimates and p-values for the best selected log(α-klotho) model............... 38

Table 14. Selected variables for different model selection criteria and k-fold cross-validation for log(eGFR) model 39

Table 15. Comparisons of model selection criteria for log (eGFR) model 40

Table 16. Parameter estimates and p-values for the best selected log (eGFR) model.................. 41

Table 17. Selected variables for different model selection criteria and k-fold cross-validation for log(serum uric acid) model 42

Table 18. Comparisons of model selection criteria for log (serum uric acid) model 42

Table 19. Parameter estimates and p-values for the best selected log (serum uric acid) model... 44

Table 20. VIF of log (α-klotho) model 45

Table 21. VIF of log (eGFR) model 45

Table 22. VIF of log (serum uric acid) model 46

Table 23. Cut point formulas and values for the three models 52

Table 24. Comparisons of model selection criteria for three models 55

Table 25. Parameter estimates and p-values for the log(ACR) model with three biomarkers 58

Table 26. Parameter estimates and p-values for the log(ACR) model with biomarkers and other variables 59

Table 27. Characteristics of children with diabetic nephropathy by ACR normal and abnormal groups 64

Table 28. Characteristics of children with diabetic nephropathy by ACR quartiles 65

Table 29. Characteristics of children with diabetic nephropathy by low HbA1c and high HbA1c groups 66

Table 30. Correlations between Hemoglobin A1c and physical characteristics, glycemic control, ... 67

Table 31. Correlations between ACR and physical characteristics, glycemic control, urine-related, serum related variables adjusting for age (Spearman correlation).................................. 68

Table 32. Correlations between estimated GFR and physical characteristics, other potential biomarkers adjusting for age (Spearman correlation) 69

Table 33. Model selection criteria with or without central obesity and waist% for eGFR model 70

Table 34. Model selection criteria without central obesity and waist% for non-missing dataset for eGFR model 70

Table 35. Model selection criteria with or without central obesity and waist% for serum uric acid model.. 71

Table 36. Model selection criteria without central obesity and waist% for non-missing dataset for serum uric acid model 72

Table 37. VIF of log (ACR) model with biomarkers and other variables 79

LIST OF FIGURES

Figure 1. Distributions of dependent variables and transformations 28

Figure 2. Scatter plot of residuals by predicted values for log (α-klotho) model 48

Figure 3. Scatter plot of residuals by predicted values for log (eGFR) model 48

Figure 4. Scatter plot of residuals by predicted values for log (serum uric acid) model 49

Figure 5. QQ plot of residuals for log (α-klotho) model 50

Figure 6. QQ plot of residuals for log (eGFR) model.. 50

Figure 7. QQ plot of residuals for log (serum uric acid) model 51

Figure 8. Plots of problematic points for log (α-klotho) model... 53

Figure 9. Plots of problematic points for log (eGFR) model... 54

Figure 10. Plots of problematic points for log (serum uric acid) model...................................... 55

Figure 11. Distribution of ACR and the log transformation.. 56

Figure 12. Fit criteria for the best selected log (α-klotho) model 69

Figure 13. Fit criteria for the best selected log (eGFR) model 71

Figure 14. Fit criteria for the best selected log (serum uric acid) model 72

Figure 15. Regression diagnostics for log (α-klotho) model 73

Figure 16. Residual analysis for log (α-klotho) model 74

Figure 17. Regression diagnostics for log (eGFR) model 75

Figure 18. Residual analysis for log (eGFR) model 76

Figure 19. Regression diagnostics for log (serum uric acid) model 77

Figure 20. Residual analysis for log (serum uric acid) model 78

Figure 21. Fit criteria for the log (ACR) model with biomarkers and other variables 79

Figure 22. Scatter plot of residuals by predicted values for log (ACR) model 80

Figure 23. QQ plot of residuals for log (ACR) model.. 81

Figure 24. Plots of problematic points for log (ACR) model 82

Figure 25. Regression diagnostics for log (ACR) model.. 83

Figure 26. Residual analysis for log (ACR) model... 84

PREFACE

I appreciate the help of my academic and thesis advisor, chair of my thesis committee, Dr. Vincent C. Arena, for providing me with an environment to conduct the research that I like. Thank you for your help through my entire master career, both in study and thesis. Hope you enjoy your retired life.

I would like to thank Dr. Ingrid M. Libman, one of my thesis committee members, for her useful suggestions in handling data and constructing models.

I also want to thank Dr. Ada O. Youk, one of my thesis committed members, for the excellent teaching of her applied regression analysis course. This course helped me a lot when I wrote my thesis.

The study and data source was from Dr. Pedro A. Pagán Banchs, a fellow in Endocrinology at Children’s Hospital of Pittsburgh. I would like to thank for him for providing me the opportunity to work on his study. I also appreciate his help during my thesis research timeline.

I would like to say thank you for all the faculty members, staff and friends who helped me in my master career. In addition, I want to thank for my parents for their care and financial supports during my whole master career. 感谢我的祖母从小到大对我无微不至的关爱，没有您就没有今天的我。愿您在天堂里过得快乐。

1.0 INTRODUCTION

1.1 TYPE 1 DIABETES MELLITUS

Diabetes Mellitus is a chronic disease which is caused by low insulin production level causing high blood sugar levels. Symptoms of Diabetes Mellitus include increased urination, increased hunger, increased thirst, and decreased weight. Diabetes Mellitus is the seventh most common disease both in the United States and in the world [1]. Currently, 29.1 million people in the United State have Diabetes Mellitus. That is 1 out of 11 people. The risk of death for adults with Diabetes Mellitus is 50% higher than in adults without Diabetes Mellitus [2]. Diabetes Mellitus can also cause many complications, including blindness, kidney diseases, heart attack and stroke [3]. T1D is a very common type of Diabetes Mellitus in the world. 5%-10% of the total cases of Diabetes Mellitus worldwide are T1D patients [4]. Currently, the incidence rate of T1D increases by about 3% per year [5].

Type 1 Diabetes Mellitus (T1D) is a type of autoimmune disease caused by genetic defect and environmental factors and may result in the destruction of beta cells. T1D occurs when the body cannot produce enough insulin and may develop for people at any age. Generally, children or young adults are at a high risk for T1D [2]. People with Type I Diabetes Mellitus usually depend on exogenous insulin throughout their remaining lifetime [6]. Currently, there is no known method to prevent and cure Type I Diabetes Mellitus The mechanism and pathology of

Type I Diabetes Mellitus is a hot research topic [9]. As a result, we need to pay more attention to the prevention and diagnosis of T1D.

1.2 DIABETIC NEPHROPATHY

Diabetic nephropathy (DN) is defined as persistent proteinuria > 500mg/24h or albuminuria > 300mg/24h [10]. Diabetic nephropathy is the leading cause of mortality in Type I Diabetes Mellitus [11]. According to the Finnish Diabetic Nephropathy Study, individuals who had T1D but without DN had an equivalent mortality compared to the general Finnish population. In contrast, individuals who had T1D and DN have much higher mortality compared to the general Finnish population [12]. Hence, it is very important to emphasize the need for prevention and early identification of DN. Recent screening guidelines from the International Society for Pediatric and Adolescent Diabetes (ISPAD) suggest that youth with T1D should be screened for DN from age 10 with 2-5 years of diabetic duration years. If puberty onset occurs earlier than age 10, young children should be screened at the onset of puberty. Screening should be performed annually [13]. Currently, albuminuria is the gold standard for screening and predicting the progression of DN in children with T1D [14]. However, it has been shown that the gold standard method lacked sensitivity for early pathological manifestations of the disease [15]. Therefore, we need to identify potential biomarkers that relate to DN or diabetic control. The assessment of potential biomarkers can be evaluated by constructing appropriate statistical models. Estimated GFR, α-klotho and serum uric acid (SUA) were selected to study novel biomarkers based on previous work and literature review.

1.3 ALPHA-KLOTHO

Klotho is a type of protein which is responsible for transporting organism to insulin. Alpha-klotho (α-klotho) is found in the human body as a necessary co-receptor molecule for the FGF23 function and is important for phosphate handling and calcitriol regulation at the kidney level [16]. The soluble form of α-klotho is derived from a cleavage of an extracellular portion of the transmembrane α-klotho (renal α-klotho). Both renal and soluble forms of α-klotho have been found to have decreased expressions in chronic kidney disease and result in bone mineral abnormalities [17]. Recently, α-klotho was shown to have negative association with albuminuria [18]. It has also been shown to be correlated with Hemoglobin A1c [17]. However, to our knowledge, no study has examined the relationships between α-klotho and different demographic and clinical factors.

1.4 ESTIMATED GFR

Glomerular filtration rate (GFR) is often used to measure kidney function and estimated GFR (eGFR) can be obtained by using a Cystatin C-based equation [20]. A Number of formulas have been established to estimate GFR and many formulas base on the creatinine. In our study, Cystatin C-based equation was used because creatinine is not stable in the human body. Serum cystatin C is a more precise reflection of kidney function than serum creatinine levels. Renal hyperfiltration (assessed by GFR) may be the earliest abnormality of kidney function and it is associated with an increased risk of DN [21]. Also, serum Cystatin C is associated with insulin resistance in patients with type 1 diabetes, which may result in cells failing to respond normally to the insulin [22]. Currently, to our knowledge, no study has examined the relationships between eGFR and different demographic and clinical predictors.

1.5 SERUM URIC ACID

Serum uric acid (SUA) predicts vascular complications in T1D and is considered one of the risk factors of T1D. It has been shown that patients with T1D have decreased SUA. In addition, some other evidence shows that SUA is associated with vascular complications in T1D [23][24]. So SUA was involved as a novel biomarker to predict DN among patients with T1D. Also, SUA has been shown to predict the development of albuminuria in adults with T1D [25] and it is associated with reduced insulin sensitivity [26]. To our knowledge, no study has examined the relationships between serum uric acid and different demographic and clinical predictors.

1.6 GOALS

The goal of this study is to assess whether potential risk biomarkers correlate with albuminuria as measured by Albumin/Creatinine Ratio (ACR), and glycemic control as measured by Hemoglobin A1c (HbA1c). Multivariate regression modeling techniques will be used to identify models for biomarkers that are not over-optimized. Appropriate variables will be selected from the candidate variables and will be used to predict the different risk biomarkers. Multiple regression modelling will be used to build models relating demographic and clinical factors to these biomarkers. Various model selection procedures will be used to select the terms in the final models and the results compared among the candidate models.

2.0 METHODS

2.1 STUDY DESIGN AND SUBJECTS

This study is a pilot/exploratory study and includes data collected from100 children who satisfied ISPAD and ADA criteria for screening for complications in T1D. The criteria are: (1) at least 10 years old, and (2) have at least 2 years duration of Type I Diabetic. Subjects were recruited from children who were seen at the Diabetes Center at the Children's Hospital of Pittsburgh of UPMC between March 2016 to June 2016.

Initial study data included information from available electronic medical records. These data included subjects' demographical information such as age, gender, height and weight. Once eligible subjects were identified, the principle investigator contacted subjects/parents/guardian and asked them if they wanted to participate in the study. A total of 100 children with T1D were identified during the recruitment time period. A more detailed review of subjects' characteristics related to T1D and DN identified three children ineligible for this study and were subsequently excluded. Thus, there were in total 97 subjects for the study.

Informed consent was obtained for all subjects. This study was approved by the University of Pittsburgh and Children's Hospital Institutional Review Board.

2.2 INDEPENDENT VARIABLES

2.2.1 Demographic variables

- **Average systolic and diastolic blood pressure percentile**

High blood pressure was considered as one of the complications caused by Diabetes [9]. Systolic and diastolic blood pressure (SBP or DBP) were measured three times at the same clinic visit when patients came to Children's Hospital. Average SBP or DBP were used to represent the corresponding blood pressure for the subject. It is typical to use BP percentile adjusted by age, gender and other variables instead of BP because BP percentile is more appropriate to represent BP characteristics and cause less bias than raw BP [27].

- **BMI percentile**

Body Mass Index (BMI) was calculated as mass/height^2 and the unit is kg/m^2. It is widely used to represent the level of shape of the human body and adiposity. BMI percentile adjusting for age, race and gender is considered a more relevant measure than BMI when assessing children as it takes into account norms across demographics subgroups Subjects with higher BMI, usually greater than 85% BMI percentile, are recognized as overweight. Subjects whose BMI percentile are greater than 95% BMI percentile are recognized as obese.

- **Waist/height ratio and central obesity**

Waist/height ratio (WHR) is defined as the waist circumference divided by the height. It is universally used to reflect the body fat and obesity for a person. Central obesity is a dichotomous variable of waist/height ratio [28]. In this study, WHR larger than 0.5 is considered to have a tendency of central obesity and more adverse health.

- **Waist circumference percentile**

Waist circumference is often used to assess the obesity level for a person. Waist circumference percentile adjusting for age, gender and race is considered more relevant than waist circumference to predict obesity or other disease because it can reflect the actual obesity level for children of different age levels, gender, and race [29].

2.2.2 Clinical variables

- **Hemoglobin A1c**

Hemoglobin A1c (HbA1c) measures plasma glucose concentration and generally it is used to represent the glycemic control level. It has been shown that HbA1c is related to Diabetes and cognition [30]. Hypoglycemia and Hyperglycemic crisis are two complications caused by Diabetes Mellitus [9]. Therefore, HbA1c would be included in the clinical variables to examine the relationships with risk biomarkers.

- **Albumin Creatinine Ratio**

Albumin Creatinine Ratio (ACR) is defined as the albumin divided by creatinine. ACR can reflect the albuminuria level, which is the non-invasive gold standard to predict the progression of DN. Also, it is an important index which can identify proteinuria and reflect the kidney function [31].

- **Triglycerides**

Triglyceride (TG) is one of the important substances which constitutes body fat [32]. Triglyceride level is considered important in predicting obesity among people. Higher Triglycerides usually represents higher body fat and thus a higher risk of obesity-related disease such as heart disease and Diabetes Mellitus [33].

- **Total cholesterol**

Cholesterol is a type of sterol and it is one of the essential components of the cell membrane. The functions of Cholesterol are to build animal membranes and to synthesize steroid hormones. It is a risk factor of cardiovascular diseases and higher cholesterol represents a higher risk of cardiovascular diseases [34].

- **HDL and LDL**

High-density lipoproteins (HDL) is one of five major lipoproteins. A low-level HDL is recognized as a risk factor of cardiovascular diseases and has close relationship with cholesterol [34]. Low-density lipoproteins (LDL) is also one of five major lipoproteins and it is well-known as a risk factor of cardiovascular diseases [35]. In addition, high blood LDL cholesterol is one of the complications caused by Diabetes Mellitus [9]. Hence, Cholesterol, HDL and LDL will be included in the model construction to see whether they are significant predictors for three novel biomarkers.

2.3 DATA MANAGEMENT

All study data from the electronic medical records was entered into an Excel by the study investigator. Upon receiving the Excel file containing the raw data, I was responsible for all of the project data management and analysis. This included a comprehensive review of all raw data, cleaning of the dataset and performing all statistical analysis including modelling and graphical analysis.

My initial step was to review the entire dataset for any errors, outliers and missing data that were not consistent with the data value specifications. This was followed by logic checks between variables within a subject. This included verifying that date sequences and combinations of variables made logical sense. All issues with the data were communicated to the study investigator and updated data files were obtained. For all transformed and calculated measures in the data set, I verified their accuracy. Prior to any analyses, I examined all variables for their distributional characteristics to assure that the statistical assumptions of the proposed analyses were met.

2.4 STATISTICAL ANALYSIS

The following parts are a brief description of correlation methods, model selection methods, model selection criteria, and problematic points methods. More details can be found in Regression analysis by Example, 5th edition [36].

2.4.1 Correlation

There are two types of correlations that are most commonly used: Pearson correlation and Spearman correlation.

- **Pearson correlation**

Pearson correlation coefficient is a parametric method to measure the linear correlation between two variables. Pearson correlation coefficient is calculated as the covariance of two variables divided by the standard deviation of each variable [35]. It ranges between -1 and 1. Pearson correlation coefficients equal to -1 or 1 indicate a perfect linear relationship between two variables. Usually, Pearson correlation is used when two variables are continuous and bivariate normally distributed.

- **Spearman correlation**

Spearman correlation coefficient is a nonparametric method to measure the rank of two variables. Spearman correlation evaluates whether the monotonic relationship exists between two variables. If Spearman correlation coefficient equals to -1 or 1, a perfect monotonic relationship is present for the two variables. Spearman correlation is appropriate for both normal or non-normal data, continuous or order categorical variables.

In our study, because some variables are not normally-distributed and ordered categorical variables are included, Spearman correlations were used to estimate associations between biomarkers and predictors.

2.4.2 Univariable regression

Univariable regression is a regression method which only includes one regressor at a time. In our study, each variable was fitted in a univariable regression model for each potential biomarker. P-values and R-square were displayed to show how well each univariable regression model performed.

2.4.3 Multiple regression

Multiple regression is an extension to the univariable regression model. A multiple regression model can include several independent variables in one model. A multiple regression model is usually a more precise model than a univariable model because it involves more predictors and can predict the outcome better. However, it is difficult to choose the best model and assess the overall fit of multiple regression model. Some model selection criteria and methods need to be assessed to evaluate the fit of multiple regression models.

2.4.4 Model selection methods

Generally, there are three types of model selection methods. They are: backwards elimination, forwards selection, and stepwise regression.

- **Backwards elimination**

Backwards elimination is one of the model selection methods which first includes all independent variables. Based on model selection criterion, the variable with the worst value of selection criterion is removed from the model. Then the model is reestimated using remaining variables and then the second worst variable is removed. These steps are repeated until the stopping rule is reached and no other variable can be removed from the model.

- **Forwards selection**

Forwards selection is one of the model selection methods which starts the model with no predictors. Based on the model selection criterion, the first variable with the best value of selection criterion is added into the model. Then the second variable with the best value of selection criterion in the remaining variable list is added into the model. These steps are repeated until the stopping rule is reached and no other variables can be added into the model.

- **Stepwise regression**

Stepwise regression is a modified method of forwards selection or backwards elimination. It allows variables to enter and leave the model at each step. The model selection criterion is computed at each step of the forwards selection or backward elimination. The variable with the worst value of selection criterion is removed from the model and the model is refitted with the remaining variables. The iteration is repeated until the stopping rule is reached and no other variable can be added or removed from the model.

In our study, the backwards elimination method was used instead of forwards selection because the maximum model could be constructed using backwards elimination. Moreover, backwards elimination can handle the collinearity issue better than forwards selection.

2.4.5 Model selection criteria

There are many model selection criteria that are commonly used in the model selection. In our study, R-square or adjusted R-square, Akaike Information Criteria (AIC), Bayes Information Criteria (BIC), Mallow’s Cp, predicted residual error sum of squares (PRESS) or cross-validation predicted residual error sum of squares (CV PRESS) were used as the model selection criteria.

- **R-square and adjusted R-square**

R-square is the squared multiple correlation coefficient. It can be interpreted as the proportion of variability of the dependent variable that can be explained by the predictive variables. The R-square increases when more variables are added into the model. It will not decrease as the number of variables increase. Large R-square close to 1 indicates the model has a good fit.

The adjusted R-square is an approach to solve the condition that R-square increases when other variables are added into the model and automatically increase R-square. This form of R-square adjusts for the number of predictors and are always less than or equal to R-square.

- **Akaike Information Criteria (AIC)**

Akaike Information Criteria (AIC) is one of the model selection criteria which aims at balancing the accuracy of the model and the simplicity of the model. Usually with the increase of predictors, the model has a better fit and accuracy. However, overfitting may occur as the number of predictors increase. Thus, it is necessary to find a balance between accuracy and simplicity. The model with smaller AIC is preferred. Also, two models with the difference of AIC no more than 2 are considered equivalent.

- **Bayes Information Criteria (BIC)**

Bayes Information Criteria (BIC) is the modified version of AIC. The difference between AIC and BIC is the severity of penalty for the number of predictors. The BIC method has a more severe penalty. As a result, the BIC method can better control the overfitting issue. The model with smaller BIC is preferred and the value of BIC is always less than AIC.

- **Mallow's Cp**

Mallow's Cp is one of the model selection criteria used to estimate the standard total mean squared error of prediction. The expectation of Cp is the number of predictors p when the model is unbiased. Thus, the model with the value of Cp closed to the number of predictors is considered a good model. The model with a small value of Mallow's Cp indicates a good precision in predicting the future outcomes.

- **Predicted residual error sum of squares (PRESS)**

Predicted residual error sum of squares (PRESS) is the sum of the residual square. It is used to assess the fit and accuracy of a model. PRESS can measure how well the model can predict for new observations [37]. Smaller PRESS values indicate better model structures. CV PRESS is the cross-validation predicted residual error sum of squares. The PRESS values are calculated in each step of cross-validation.

2.4.6 Model validation techniques

After fitting the appropriate model, different model validation technique can be used to validate the model and test how well the model predicts for new observations. Different model validation techniques are described below:

- **Least Absolute Shrinkage and Selection Operator (LASSO)**

Least Absolute Shrinkage and Selection Operator (LASSO) is a type of least squares regression. LASSO finds the values of regression coefficients which minimize the mean residual sum of squares, where the sum of absolute values of coefficients is constrained to less than or equal to a constant t. Since the t is usually a small value, the regression coefficients are usually close to zero. LASSO can improve the accuracy of selected models and it can be used for many statistical models such as regression models, generalized linear models and proportional hazard models. However, since many coefficients are small and close to zero, sometimes it is hard to interpret the model using LASSO.

- **Bootstrapping**

Bootstrapping is a model validation technique which can be used to evaluate the models and parameter estimates without making assumptions for the distributions. It can be used when the sample size is small and overfitting issue exists. For bootstrapping, B repeated samples of size n are drawn with replacement. Then the analysis is repeated based on each of the B datasets. The parameter estimates are based on the average of the B bootstrap samples. It is easy to derive the point estimates and confidence interval of parameters using bootstrapping. However, although the bootstrapping can be used to small simple size, it could still cause some bias.

- **Cross-validation**

Cross-validation is a model validation method to evaluate how well the predictive model performs and how well they can be applied to other observations [38]. If the whole dataset is used to construct the model and predict the outcomes, no other data would be used to examine how well the model will be predictive for new data,

The common cross-validation method splits data into two parts: training set and testing set. Training set is used to fit the model and predict the outcomes, Testing set is used to test how well the predictive model performs for another dataset.

Exhaustive and Non-exhaustive cross-validation are two categories of this method. Exhaustive cross-validation tests all possible ways for splitting the dataset. There are two major methods for exhaustive cross-validation: leave-one-out cross-validation and leave-p-out cross-validation.

Leave-p-out cross-validation puts p observations into the testing set and the remaining n-p observations into the training set. Each combinations of p from the sample sizes n are chosen exactly one time as the testing set and this method chooses all possible combinations of p from n. As a result, this method is extremely time-consuming, especially when n is large.

Leave-one-out cross-validation is a special case of leave-p-out cross-validation when p=1 and it is similar to jackknifing. Each observation is chosen as the testing set for exactly one time so there are in total n possible combinations for the selection of testing set. This cross-validation method is not time-consuming for our study because the sample size of our study is relatively small.

Non-exhaustive cross-validation do not test all possible ways for splitting the dataset. There are many non-exhaustive cross-validation methods, which are introduced as follow.

K-fold cross-validation is one commonly-used exhaustive cross-validation method. In this method, the dataset is partitioned into k subsets of equal size. Each subset is treated as the testing set for one time and the remaining k-1 subsets are treated as the training set. For example, if k=5, the original dataset is partitioned into 5 subgroups. At each time, 1 of 5 subgroups will be selected as the testing set and the other 4 groups will be the training set. The iterations will be repeated five times until each of five subgroups has been chosen to be the test set for exactly one time. Different k values will affect the results of cross-validation. If k=n, the n-fold cross-validation is the same method as leave-one-out cross-validation. Hence, the choice of k is important and it should be assessed when we are doing the model selection.

Hold out method is the simplest cross-validation method. It randomly splits the data into training set and testing set. But it runs just for a single time. This method lacks accuracy because it only uses one training set and one testing set to predict and validate the model. The results could be different based on the choice of training set and testing set.

Monte Carlo cross-validation randomly splits the data into training set and testing set. The Monte Carlo cross-validation runs multiple time based on different splits, which can largely reduce the bias caused by choice of training set and testing set. However, some observations in the dataset may not be selected in the testing set even one time. Some observations could be selected in the testing set for more than one time.

In our study, since we want each observation to be treated as the testing observation for one time and models that are easy to construct and interpret, we will use only k-fold cross-validation methods as the model validation techniques based on different choice of k.

The feature of the study is to use k-fold cross-validation methods to choose appropriate models for the three potential biomarkers. CV PRESS is used as the selection rule in our study. Different model selection criteria are used to compare and select the final models for each choice of k. In addition, selection results for different choices of k for k-fold cross-validation are compared in order to select the best cross-validation method for three models.

2.4.7 Problematic points checking

There are many types of problematic points that are commonly checked after the construction of models. In our study, outliers, leverage, and influence were checked to see whether the model had any problematic points.

- **Outliers**

An Outlier is an observation which is far more different than other observations [39]. Outliers can be assessed by graphical residual analysis such as histogram, stem and leaf, box plot or scatter plot. In our study, scatter plots of studentized residuals were used to find outliers. Any values larger than ± 2 were considered an outlier, which correspond to approximately 5% of residuals when residuals were normally distributed.

- **Leverage**

In the multiple regression, assuming that the model is $Y=X\beta + \varepsilon$, $X(X'X)^{-1}X'$ is called the Hat matrix. The i^{th} diagonal element of the Hat matrix is called the leverage of i^{th} observation. High leverage points usually represent that those points have undue influence on the model. In our study, the scatter plots of studentized residuals versus leverage were used to find any high leverage points.

- **Influence**

Influence measures how much the regression coefficients change when ith observation is deleted from the model. If the deletion of ith observation causes a large change for the coefficients, that observation can be viewed as the influential point. There are many methods which can be used to detect the influential points. In our study, Cook’s Distance and DFFITS were used to find influential points.

Cook’s Distance is a squared distance between estimated parameters for all the dataset and estimated parameters for the dataset when ith observation is dropped. Higher Cook’s Distance values indicate that the deletion of that observation may have a large influence on the regression coefficients. In our study, Cook’s Distance was assessed by a plot of Cook’s Distance versus the observation No.

DFFITS is another way to detect the influential points. It also measures how much the regression coefficients change when ith observation is deleted from the model. It is similar with studentized residual. A high DFFITS value or a low DFFITS value indicates a high influential point. In our study, DFFITS was assessed by a plot of DFFITS versus the observation No.

2.4.8 Software

All parts of statistical analysis were done by using SAS 9.4 (SAS Institute, Cary, NC).

3.0 RESULTS

3.1 DESCRIPTIVE CHARACTERISITICS

3.1.1 Continuous variables

The descriptive analysis of continuous variables was based on 97 eligible subjects and 17 selected continuous variables (Table 1). The age of subjects was 15.8 ± 2.9 and ranged between 10.0 and 23.6. The mean diabetic duration year was 7 years and ranged between 2.1 to 17.6. The median BMI percentile was 75.9% and larger than the mean BMI percentile (65.7%), which indicated that over half of the children were overweight. Other descriptive characteristics of continuous variables are shown in Table 1. Table 1 also shows that there was missing data for several variables.

Table 1. Descriptive statistics for continuous variables

Variable	N	Mean	Std Dev	25th Pctl	Median	75th Pctl	Min	Max
Age (years)	97	15.82	2.89	13.87	15.44	17.73	10.04	23.57
Diabetes Duration (yrs)	97	7.02	3.81	4.18	6.34	9.65	2.08	17.63
Average SBP (%)	90	54.43	23.88	36.91	53.81	74.03	1.25	94.91
Average DBP (%)	90	69.16	16.88	59.91	71.16	80.41	20.74	92.64
BMI (%)	90	66.01	29.37	49.28	75.88	88.97	0.06	99.05
Waist/height ratio	74	0.48	0.06	0.44	0.47	0.53	0.36	0.66
Growth velocity (cm/yr)	96	2.77	2.68	0.50	1.60	4.90	0.00	12.00
Hemoglobin A1c (%)	97	8.08	1.31	7.30	8.00	8.60	5.51	11.94
ACR (mg/g)	88	16.48	18.83	5.50	8.95	19.30	0.00	90.00
α-klotho (pg/mL)	79	1301.88	569.01	871.44	1204.71	1555.23	439.65	3334.83
eGFR (ml/kg/1.73m^2)	86	99.85	20.79	87.05	96.49	109.75	61.35	179.25
Serum Uric Acid (mg/dL)	93	3.79	0.95	3.10	3.70	4.40	1.90	6.30
Creatinine (mg/dL)	93	1.65	9.40	0.60	0.70	0.80	0.40	91.30
Triglycerides (mg/dL)	97	103.12	69.76	62.00	90.00	120.00	23.00	492.00
Total cholesterol (mg/dL)	97	165.90	27.40	147.00	163.00	184.00	72.00	245.00
LDL (mg/dL)	97	90.15	23.91	76.00	89.00	102.00	11.00	150.00
HDL (mg/dL)	97	56.95	11.67	49.00	56.00	65.00	34.00	86.00

3.1.2 Categorical variables

The descriptive analysis of categorical variables was based on 97 eligible subjects and 6 selected continuous variables (Table 2). For race, white and African American (AA) are the only two races that are being considered. Central Obesity was defined as waist/height ratio ≥ 0.5. Waist circumference was coded as 0, 1, 2, 3, 4, 5, 6, corresponding to standardized percentiles groups of 0%~10%, 10%~25%, 25%~50%, 50%~75%, 75%~85%, 85%~90% and 90%~100%. A dichotomous variable for Hemoglobin A1c was created and the cut point was 7.5 based on previous literature. Values higher than 7.5 indicated less glycemic control and values less than 7.5 indicated more glycemic control. ACR was divided into two groups. Values less than or equal to 30 were classified into the normal group and values larger than 30 were classified into the abnormal group. Most subjects were white and had no central obesity. Few children had a large waist circumference after adjusting for their age, sex, and race. Subjects whose HbA1c ≥ 7.5 were nearly two times more than whose HbA1c < 7.5. 74 out of 88 subjects had normal ACR.

Table 2. Descriptive statistics for categorical variables

Variable	Frequency	Percent
Sex	97	
Male	50	51.6
Female	47	48.4
Race	95	
White	86	90.6
AA	9	9.4
Central Obesity	75	
Yes	25	33.8
No	49	66.2
Waist Percentage	73	
0%	6	8.3
10%	11	15.3
25%	12	16.7
50%	25	34.7
75%	9	12.5
85%	3	4.2
90%	6	8.3
Hemoglobin A1c (%)	97	
< 7.5	32	33.0
≥ 7.5	65	67.0
ACR (mg/g)	89	
Normal	74	84.1
Abnormal	14	15.9

3.2 CORRELATIONS

3.2.1 α-klotho and independent variables

Correlations between soluble α-klotho (N=79) and other variables are shown in Table 4. Statistically significant negative correlations are noted between soluble α-klotho and age (r=-0.32, p=0.004), diabetes durations (r=-0.45, p<0.0001), waist/height ratio (r=-0.38, p=0.004) and HbA1c (r=-0.30, p=0.007). A significant positive correlation is noted between soluble α-klotho and growth velocity (r=0.24, p=0.04).

Table 3. Correlations between α-klotho and other independent variables

Variable	R	p-value	N
Age (years)	-0.32	0.004	79
Diabetes Duration(yrs)	-0.45	<0.0001	79
Average SBP (%)	0.008	0.95	72
Average DBP (%)	-0.08	0.49	72
BMI %	-0.06	0.59	72
Waist/height ratio	-0.38	0.004	57
Growth velocity (cm/yrs)	0.24	0.04	78
Hemoglobin A1c (%)	-0.30	0.007	79
ACR (mg/g)	0.02	0.89	75
Creatinine (mg/dL)	0.06	0.61	78
Triglycerides (mg/dL)	-0.12	0.31	79
Total cholesterol (mg/dL)	-0.11	0.34	79
LDL (mg/dL)	-0.06	0.59	79
HDL (mg/dL)	-0.10	0.38	79

3.2.2 Estimated GFR and independent variables

Correlations were examined between eGFR (N=86) and other independent (Table 3). Significant positive correlations are found between eGFR and average DBP (r=0.30, p=0.007) and waist/height ratio (r=0.35, p=0.004). Significant negative correlations are found between eGFR and growth velocity (r=-0.29, p=0.007) and creatinine (r=-0.34, p=0.001).

Table 4. Correlations between eGFR and other independent variables

Variable	R	p-value	N
Age (years)	0.10	0.37	86
Diabetes Duration(yrs)	-0.14	0.18	86
Average SBP (%)	0.07	0.52	81
Average DBP (%)	0.30	0.007	81
BMI %	0.07	0.56	81
waist/height ratio	0.35	0.004	66
Growth velocity (cmy/yr)	-0.29	0.007	85
Hemoglobin A1c (%)	0.19	0.08	86
ACR (mg/g)	0.19	0.10	78
Creatinine (mg/dL)	-0.34	0.001	86
Triglycerides (mg/dL)	0.10	0.34	86
Total cholesterol (mg/dL)	0.17	0.11	86
LDL (mg/dL)	0.12	0.28	86
HDL (mg/dL)	0.13	0.23	86

3.2.3 Serum uric acid and independent variables

Correlations between serum uric acid (N=93) and other variables are shown below (Table 5). A significant negative correlation between serum uric acid and HbA1c (r=-0.24, p=0.02) is noted. Significant positive correlations between serum uric acid and age (r=0.27, p=0.009) and creatinine (r=0.47, p<0.0001).

Table 5. Correlations between serum uric acid and other independent variables

Variable	R	p-value	N
Age (years)	0.27	0.009	93
Diabetes Duration(yrs)	0.01	0.90	93
Average SBP (%)	0.05	0.63	86
Average DBP (%)	-0.14	0.20	86
BMI %	0.13	0.22	86
Waist/height ratio	0.02	0.85	70
Growth velocity (cm/yrs)	-0.07	0.50	92
Hemoglobin A1c (%)	-0.24	0.02	93
ACR (mg/g)	-0.06	0.58	85
Creatinine (mg/dL)	0.47	<0.0001	93
Triglycerides (mg/dL)	0.02	0.84	93
Total cholesterol (mg/dL)	-0.17	0.11	93
LDL (mg/dL)	-0.16	0.13	93
HDL (mg/dL)	-0.09	0.40	93

3.3 TRANSFORMATIONS OF DEPENDENT VARIABLES

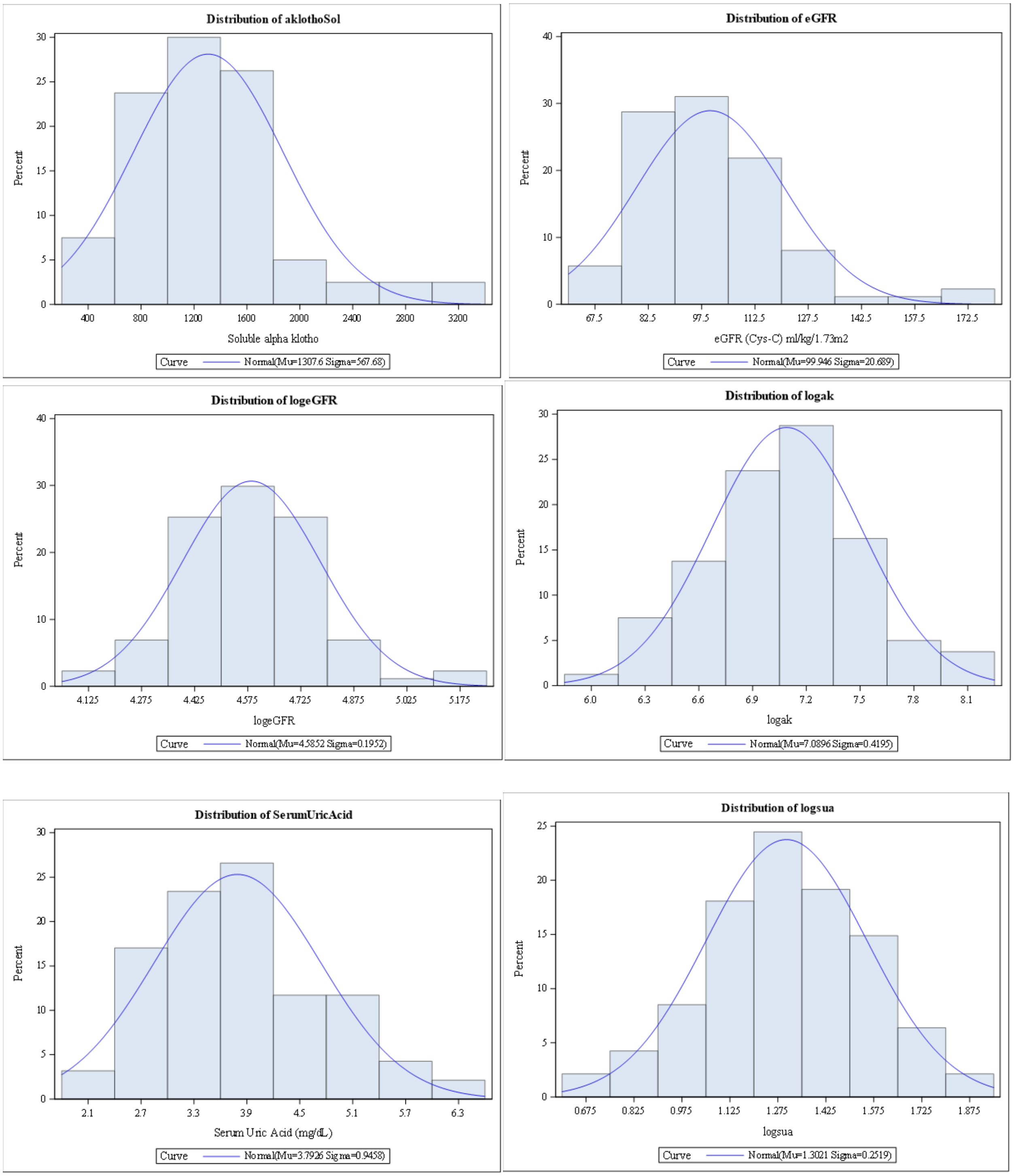


Figure 1. Distributions of dependent variables and transformations

Before performing univariable analysis, the normality of dependent variables eGFR, α-klotho, and serum uric acid were checked. Figure 1 displays histogram analysis of normality. The left three graphs are histograms of three dependent variables without transformation. All of them are positive-skewed showing some departures from normality may exist. The right three graphs are histograms of three dependent variables with log transformation. After log transformation, all three dependent variables seemed to be normally distributed. Hence, log transformation of these three dependent variables were used in the modelling.

3.4 UNIVARIABLE REGRESSION

Univariable regression models between three potential biomarkers and each independent variable were built to assess whether there was a statistically significant association between each biomarker and each independent variable.

3.4.1 Log (α-klotho)

Table 6 shows the univariable regression results between log (α-klotho) and each independent variable. Log (α-klotho) was significantly associated with age (p=0.004, R^2= 0.104), diabetes duration (p<0.0001, R^2= 0.208), waist/height ratio (p=0.007, R^2=0.126), growth velocity (p=0.027, R^2=0.063), HbA1c (p=0.009, R^2=0.087), Creatinine (p=0.012, R^2=0.080), central obesity (p=0.051, R^2=0.069), and HbA1c group (p=0.037, R^2=0.055).

Table 6. Univariable regression models for log α-klotho

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value	R-Square
Age (years)	1	-0.046	0.015	-2.99	0.004	0.104
Diabetes Duration(yrs)	1	-0.049	0.011	-4.49	<.0001	0.208
Average SBP (%)	1	0.001	0.002	0.47	0.643	0.003
Average DBP (%)	1	-0.003	0.003	-1.09	0.280	0.017
BMI %	1	0.001	0.002	0.76	0.452	0.008
Waist/height ratio	1	-2.374	0.851	-2.79	0.007	0.126
Growth velocity (cm/yrs)	1	0.038	0.017	2.25	0.027	0.063
Hemoglobin A1c (%)	1	-0.094	0.035	-2.70	0.009	0.087
ACR (mg/g)	1	0.001	0.002	0.44	0.660	0.003
Creatinine (mg/dL)	1	0.012	0.005	2.57	0.012	0.080
Triglycerides (mg/dL)	1	-0.001	0.001	-1.15	0.254	0.017
Total cholesterol (mg/dL)	1	-0.002	0.002	-1.29	0.199	0.021
LDL (mg/dL)	1	-0.002	0.002	-1.09	0.278	0.015
HDL (mg/dL)	1	0.002	0.002	1.00	0.320	0.012
Sex	1	0.128	0.094	1.36	0.177	0.024
Race	1	0.105	0.167	0.63	0.532	0.005
Central Obesity	1	-0.228	0.115	-1.99	0.051	0.069
Waist percentage	1	-0.034	0.035	-0.97	0.336	0.018
HbA1c group	1	-0.207	0.097	-2.12	0.038	0.055
ACR group	1	0.034	0.126	0.27	0.789	0.001

3.4.2 Log (eGFR)

Table 7 displays univariable regression results between log (eGFR) and each independent variable. Log (eGFR) was significantly associated with the average DBP (p=0.019, R^2=0.068), waist/height ratio (p=0.027, R^2=0.074), growth velocity (p=0.016, R^2=0.068), HbA1c (p=0.006,

R^2=0.087), sex (p=0.027, R^2=0.057), race (p=0.041, R^2=0.050), central obesity (p=0.044, R^2=0.062), and HbA1c group (p=0.042, R^2=0.048).

Table 7. Univariable regression models for log eGFR

Variable	DF	Parameter Estimate	Standard Error	t Value	p-value	R-Square
Age (years)	1	0.005	0.008	0.71	0.483	0.006
Diabetes Duration(yrs)	1	-0.006	0.006	-1.07	0.286	0.014
Average SBP (%)	1	0.001	0.001	0.55	0.585	0.004
Average DBP (%)	1	0.003	0.001	2.39	0.019	0.068
BMI %	1	0.0003	0.001	0.45	0.655	0.003
Waist/height ratio	1	1.013	0.449	2.26	0.027	0.074
Growth velocity (cm/yrs)	1	-0.019	0.008	-2.45	0.016	0.068
Hemoglobin A1c (%)	1	0.044	0.016	2.82	0.006	0.087
ACR (mg/g)	1	0.001	0.001	0.90	0.372	0.011
Creatinine (mg/dL)	1	-0.002	0.002	-1.05	0.298	0.013
Triglycerides (mg/dL)	1	0.0003	0.0002	0.92	0.361	0.010
Total cholesterol (mg/dL)	1	0.001	0.001	1.48	0.142	0.026
LDL (mg/dL)	1	0.001	0.001	1.48	0.143	0.025
HDL (mg/dL)	1	0.002	0.002	1.00	0.320	0.012
Sex	1	0.093	0.041	2.25	0.027	0.057
Race	1	-0.168	0.081	-2.08	0.041	0.050
Central Obesity	1	0.112	0.054	2.06	0.044	0.062
Waist percentage	1	0.017	0.018	0.97	0.334	0.015
HbA1c group	1	0.090	0.044	2.06	0.042	0.048
ACR group	1	0.079	0.057	1.40	0.167	0.0250

3.4.3 Log (serum uric acid)

Table 8 shows univariable regression results between log (serum uric acid) and each independent variable. Log (serum uric acid) was significantly associated with age (p=0.010, R^2= 0.071), HbA1c (p=0.023, R^2=0.056), sex (p=0.019, R^2=0.059) and HbA1c group (p=0.001, R^2=0.123).

Table 8. Univariable regression models for log serum uric acid

Variable	DF	Parameter Estimate	Standard Error	t Value	p-value|	R-Square
Age (years)	1	0.023	0.009	2.63	0.010	0.071
Diabetes Duration(yrs)	1	0.002	0.007	0.24	0.811	0.001
Average SBP (%)	1	0.0001	0.001	0.07	0.943	0.0001
Average DBP (%)	1	-0.002	0.002	-1.34	0.184	0.021
BMI %	1	0.001	0.001	1.08	0.284	0.014
Waist/height ratio	1	-0.086	0.566	-0.15	0.880	0.0003
Growth velocity (cm/yrs)	1	-0.006	0.010	-0.62	0.538	0.004
Hemoglobin A1c (%)	1	-0.045	0.019	-2.32	0.023	0.056
ACR (mg/g)	1	-0.001	0.002	-0.68	0.499	0.006
Creatinine (mg/dL)	1	0.001	0.003	0.50	0.620	0.003
Triglycerides (mg/dL)	1	-0.00002	0.0004	-0.05	0.957	0.0001
Total cholesterol (mg/dL)	1	-0.001	0.001	-1.36	0.176	0.020
LDL (mg/dL)	1	-0.002	0.001	-1.57	0.120	0.026
HDL (mg/dL)	1	-0.002	0.002	-0.73	0.464	0.006
Sex	1	-0.122	0.051	-2.39	0.019	0.059
Race	1	0.023	0.092	0.25	0.801	0.001
Central Obesity	1	0.020	0.070	0.29	0.774	0.001
Waist percentile	1	0.011	0.021	0.55	0.587	0.005
HbA1c group	1	-0.188	0.052	-3.58	0.001	0.123
ACR group	1	-0.071	0.076	-0.93	0.353	0.010

3.5 MULTIPLE REGRESSION

3.5.1 Variable selection criteria

Before fitting multiple regression models, correlations between each independent variable were checked to avoid any highly-correlated variables, which may cause multicollinearity issues. LDL and total cholesterol were highly-correlated (r=0.83). As a result, LDL and total cholesterol were not considered in the same multiple regression model. In our study, total cholesterol was included in the variable list. In addition, central obesity was created based on waist/height ratio and it is commonly used in diabetic-related research as a measure of risk. Because we were interested in determining the relationship between novel biomarkers and (1) ACR, the current gold standard to detect early DN and (2) HbA1c, which measured diabetes control, ACR and HbA1c included as continuous variables were candidate predictors. Hence, the maximum model contained age, diabetic duration, average SBP percentile, average DBP percentile, BMI percentile, growth velocity, HbA1c, ACR, Creatinine, Triglycerides, total cholesterol, HDL, sex, race, central obesity and waist percentile.

3.5.2 Cross Validation

Cross-validation was used to select variables and assess the accuracy of predictive models and include 2-fold, 5-fold, 10-fold and n-fold. Different model selection criteria such as adjusted R-square, AIC, BIC, Mallow’s Cp and CV PRESS were estimated in order to choose the best model. Here, backward elimination was used as the model selection method and CV was used as the selection and stopping criterion. CV here stands for predicted residual sum of square with k-fold cross-validation. It was used as the model selection criterion instead of significance levels because CV can provide the fit of candidate models and their model structures by using cross-validation. Predicted residual sum of square can measure the how predictive the model is when new observations are added. Moreover, the p-value was not a good selection rule when the sample size was small.

In SAS, the maximum models were constructed based on the backwards elimination. At each step, the variable with the lowest CV PRESS value when removed from the model was dropped from the model that was constructed at the last step. The selection steps ends when the CV PRESS does not decrease any further when any other variables are removed from the model. Thus, the final model was the model with the lowest CV PRESS.

3.5.2.1 Log (α-klotho) model

Table 9 displays selected variables based on selection criteria and cross-validation methods for log (α-klotho) model. Different k-fold cross-validation methods selected different numbers of variables. For each cross-validation method, the adjusted R-square selection criterion tended to include more variables than any other selection criteria. The consistent selection results could be reached when BIC, Mallow’s Cp and CV PRESS were used as selection criteria for each cross-validation method.

Adjusted R-square would not be used as the model selection criteria here because maximizing adjusted R-square could not minimize the predicted residual sum of square and the adjusted R-square tended to reach the maximum at the early selection step. The same variables were selected based on BIC, Mallow’s Cp and CV PRESS. As a result, any one of those three model selection criteria could be chosen as the selection criterion for the log (α-klotho) model.

Table 9. Selected variables for different model selection criteria and k-fold cross-validation for log(α-klotho) model

	No CV	**2-fold CV**	**5-fold CV**	**10-fold CV**	**n-fold CV**
Adjusted R-square	HbA1c, SBP%, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%	Diabetic duration, age, SBP%, DBP%, growth velocity, ACR, Creatinine, TG, Cholesterol, HDL	Diabetic duration, age, SBP%, DBP%, BMI%, growth velocity, Creatinine, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, age, DBP%, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, SBP%, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%
AIC	HbA1c, SBP%, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%	Diabetic duration, age, SBP%, DBP%, growth velocity, ACR, Creatinine, TG, Cholesterol, HDL	Growth velocity, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, growth velocity, TG, Cholesterol, HDL, central obesity	HbA1c, SBP%, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%
BIC	HbA1c, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%	Diabetic duration, age, SBP%, DBP%, growth velocity, ACR, Creatinine, TG, Cholesterol, HDL	Growth velocity, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, growth velocity, TG, Cholesterol, HDL, central obesity	HbA1c, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%
Mallow's Cp	HbA1c, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%	Diabetic duration, age, SBP%, DBP%, growth velocity, ACR, Creatinine, TG, Cholesterol, HDL	Growth velocity, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, growth velocity, TG, Cholesterol, HDL, central obesity	HbA1c, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%
CV PRESS or PRESS	HbA1c, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%	Diabetic duration, age, SBP%, DBP%, growth velocity, ACR, Creatinine, TG, Cholesterol, HDL	Growth velocity, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, growth velocity, TG, Cholesterol, HDL, central obesity	HbA1c, growth velocity, ACR, TG, Cholesterol, HDL, central obesity, waist%

Table 10. Comparisons of model selection criteria for log (α-klotho) model

	R^2	Adjusted R^2	AIC	BIC	Mallow's Cp	CV PRESS
10-fold CV	0.3710	0.2717	-51.10	-91.87	3.18	5.01
n-fold CV	0.4236	0.2955	-51.03	-88.55	4.33	5.11

Because 2-fold cross-validation and 5-fold cross-validation methods introduced more bias for true errors and predicted errors than 10-fold and n-fold cross-validation methods, only

10 and n-fold cross validation methods were considered when selecting models. Table 10 shows values of different model selection criteria for 10-fold and n-fold cross-validation methods. Since R-square of n-fold cross-validation was larger than the 10-fold one, and there were no significant differences when comparing other model selection criteria. Thus, the n-fold cross-validation method appears to be the most appropriate method for the α-klotho model.

However, because there were missing data, especially for central obesity and waist percentile, the sample size of the selected model is small. Seventy-nine subjects had data for α-klotho, but only 45 subjects did not have missing values for variables in the maximum models and thus only 45 subjects were included in the regression analysis. Hence, models excluding variables with many missing values should be constructed and compared to the n-fold cross-validation model.

When central obesity and waist percentile were excluded from the model, the sample size increased from 45 to 66. Table 11 displays values of different model selection criteria for n-fold cross-validation model with and without central obesity and waist percentile. For the model without central obesity and waist percentile, the selected model included HbA1c, age, ACR, Triglycerides and HDL. The R-square was greatly reduced, which indicated that less variability of the dependent variable could be explained by the model. Mallow's Cp and CV PRESS for the model without central obesity and waist percentile were appreciably greater than the one with central obesity and waist percentile, which suggested that the model without central obesity and waist percentile had a poor fit and did not predict new observations well. Although the model with central obesity and waist percentile had small sample size, the model was more reasonable and predictive.

Table 11. Model selection criteria with or without central obesity and waist% for α-klotho model

N-fold CV	N	R^2	Adjusted R^2	AIC	BIC	Mallow's Cp	CV PRESS
With central obesity and waist%	45	0.4236	0.2955	-51.03	-88.55	4.33	5.11
Without central obesity and waist%	66	0.2388	0.1889	-64.21	-129.85	7.68	8.81

Table 12. Model selection criteria without central obesity and waist% for non-missing dataset for α-klotho model

N-fold CV		R^2	Adjusted R^2	AIC	BIC	Mallow's Cp	CV PRESS
Without central obesity and waist% for non-missing dataset	45	0.3010	0.2311	-50.35	-93.63	3.21	4.67

A subset analysis was performed on the dataset excluding missing values for central obesity and waist percentile. The sample size was 45, which was the same as the first selected model. But this time central obesity and waist percentile were not considered as candidate variables in the maximum model. This model contained diabetic duration, growth velocity, Creatinine and Triglycerides. Table 12 displays model selection criteria of this model. Comparing this model with the model without central obesity and waist percentile on the full dataset, R-square was largely improved. The Mallow's Cp and CV PRESS decreased from 9 to less than 5, which indicated this model had a better fit and it was a more predictive model after removing 21 subjects with missing values of central obesity and waist percentile. Comparing this model with the model with central obesity and waist percentile, the model with central obesity and waist percentile had a higher R-square. In addition, central obesity was selected in that model, which indicated it was an important predictor. Therefore, central obesity and waist percentile should be retained in the maximum model.

Thus, the best log (α-klotho) model was the one using n-fold cross-validation method, BIC or Mallow’s Cp or CV PRESS as model selection criterion, and includes central obesity and waist percentile. This method contained HbA1c, growth velocity, Triglycerides, total Cholesterol, HDL and central obesity. The parameter estimates and p-values are shown in the Table 13.

Table 13. Parameter estimates and p-values for the best selected log(α-klotho) model

Parameter	Estimate	Standard Error	p-value
Intercept	9.053	0.821	<.0001
HbA1c	-0.049	0.034	0.158
Growth velocity	0.039	0.020	0.061
ACR	0.003	0.002	0.179
Triglycerides	-0.002	0.001	0.030
Total Cholesterol	0.003	0.002	0.133
HDL	-0.015	0.006	0.010
Central obesity	-3.037	1.781	0.097
Waist%	0.084	0.066	0.210

For the best selected log (α-klotho) model, R-square = 0.424, AIC = -51.03, BIC = -88.55, Mallow’s Cp = 4.33, CV PRESS = 5.11.

3.5.2.2 Log (eGFR) model

Table 14 displays selected variables based on selection criteria and cross-validation methods for the log (eGFR) model. Except for no cross-validation, other model selections based on the cross-validation method selected the same variables no matter what selection criterion was used. This also suggested the model fitted well and all model selection criteria reached the same results.

Table 14. Selected variables for different model selection criteria and k-fold cross-validation for log(eGFR) model

	No CV	2-fold CV	5-fold CV	10-fold CV	n-fold CV
Adjusted R-square	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity	HbA1c, diabetic duration, DBP%, SBP%, ACR, Creatinine, TG, Cholesterol, sex, race, central obesity, waist%	HbA1c, age, growth velocity, ACR, Creatinine, sex, central obesity, waist%	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity, waist%	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity
AIC	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity	HbA1c, diabetic duration, DBP%, SBP%, ACR, Creatinine, TG, Cholesterol, sex, race, central obesity, waist%	HbA1c, age, growth velocity, ACR, Creatinine, sex, central obesity, waist%	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity, waist%	HbA1c, DBP%, growth velocity, Creatinine, Cholesterol, central obesity
BIC	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity	Diabetic duration, ACR, Creatinine, TG, Cholesterol, sex, race, central obesity, waist%	HbA1c, age, growth velocity, ACR, Creatinine, sex, central obesity, waist%	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity, waist%	HbA1c, DBP%, growth velocity, Creatinine, Cholesterol, central obesity
Mallow's Cp	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity	HbA1c, diabetic duration, DBP%, SBP%, ACR, Creatinine, TG, Cholesterol, sex, race, central obesity, waist%	HbA1c, age, growth velocity, ACR, Creatinine, sex, central obesity, waist%	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity, waist%	HbA1c, DBP%, growth velocity, Creatinine, Cholesterol, central obesity
CV PRESS or PRESS	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity	Diabetic duration, ACR, Creatinine, TG, Cholesterol, sex, race, central obesity, waist%	HbA1c, age, growth velocity, ACR, Creatinine, sex, central obesity, waist%	HbA1c, DBP%, growth velocity, ACR, Creatinine, Cholesterol, central obesity, waist%	HbA1c, DBP%, growth velocity, Creatinine, Cholesterol, central obesity

Because 10-fold cross-validation and n-fold cross-validation were less biased than the others, the selected model was chosen based on the comparisons of different model selection criteria (Table 15). Here, n-fold cross-validation would be used to select the best eGFR model. Compared to the two cross-validation methods, there were no appreciable difference in R-square, AIC, BIC, and CV PRESS. Looking at the p-values of selected variables, the p-values of n-fold cross-validation were all significant or close to the significance level. Moreover, the n-fold cross-validation method is more precise than the 10-fold cross-validation method. As a result, the n-fold cross-validation method was used in the eGFR model.

Table 15. Comparisons of model selection criteria for log (eGFR) model

	R^2	Adjusted R^2	AIC	BIC	Mallow's Cp	CV PRESS
10-fold CV	0.6242	0.5543	-139.14	-183.24	1.56	1.26
n-fold CV	0.6176	0.5568	-140.23	-185.62	0.18	1.24

There were many missing values for central obesity and waist percentile. The sample size of the model with central obesity and waist percentile was 52. The sample size of the model without central obesity and waist percentile was 70. Therefore, the model for 70 subjects without central obesity and waist percentile and the subset analysis for 52 without central obesity and waist percentile were constructed following the previous procedures and logic of the α-klotho model. The tables of comparisons of model selection criteria for the three models were presented in the Appendix A.

After comparisons of three eGFR models with and without central obesity and waist percentile, using similar logic and analysis as α-klotho model, the best selected model for log (eGFR) used 10-fold cross-validation method and included central obesity and waist percentile. This model contained HbA1c, DBP percentile, growth velocity, ACR, Creatinine, Cholesterol and central obesity. The parameter estimates and p-values are shown in the Table 16.

Table 16. Parameter estimates and p-values for the best selected log (eGFR) model

Parameter	Estimate	Standard Error	p-value
Intercept	4.108	0.316	<.0001
HbA1c	0.028	0.015	0.069
DBP%	0.003	0.001	0.024
Growth velocity	-0.028	0.010	0.006
ACR	0.003	0.001	0.006
Creatinine	-0.826	0.146	<.0001
Cholesterol	0.001	0.001	0.138
Central obesity	0.967	0.433	0.031

For the best selected log (eGFR) model, R-square = 0.618, AIC = -140.23, BIC = -185.62, Mallow’s Cp = 0.18, CV PRESS = 1.24.

3.5.2.3 Log (serum uric acid) model

Table 17 displays selected variables based on selection criteria and cross-validation methods for the log (serum uric acid) model. For each cross-validation method, selected variables were the same for BIC and CV PRESS. Thus, BIC or CV PRESS could be used as the selection criterion.

Table 17. Selected variables for different model selection criteria and k-fold cross-validation for log(serum uric acid) model

	No CV	**2-fold CV**	**5-fold CV**	**10-fold CV**	**n-fold CV**
Adjusted R-square	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, gender, race, central obesity, waist%	HbA1c, ACR, Creatinine, TG, Cholesterol, HDL, gender, race, central obesity, waist%	All variables	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, race, central obesity, waist%	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, gender, race, central obesity, waist%
AIC	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, ACR, Creatinine, TG, Cholesterol, HDL, gender, race, central obesity, waist%	HbA1c, Diabetic duration, age, Creatinine, TG, Cholesterol	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%
BIC	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, ACR, Creatinine, HDL, gender, central obesity, waist%	HbA1c, Diabetic duration, age, Creatinine, TG, Cholesterol	HbA1c, diabetic duration, ACR, Creatinine, TG, Cholesterol, HDL	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%
Mallow's Cp	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, ACR, Creatinine, HDL, gender, central obesity, waist%	HbA1c, Diabetic duration, age, Creatinine, TG, Cholesterol	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%
CV PRESS or PRESS	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%	HbA1c, ACR, Creatinine, HDL, gender, central obesity, waist%	HbA1c, Diabetic duration, age, Creatinine, TG, Cholesterol	HbA1c, diabetic duration, ACR, Creatinine, TG, Cholesterol, HDL	HbA1c, diabetic duration, age, ACR, Creatinine, TG, Cholesterol, HDL, central obesity, waist%

Table 18. Comparisons of model selection criteria for log (serum uric acid) model

	R^2	Adjusted R^2	AIC	BIC	Mallow's Cp	CV PRESS
10-fold CV	0.3570	0.2591	-89.63	-141.17	8.66	3.61
n-fold CV	0.4467	0.3181	-91.75	-138.40	8.15	3.48

Table 18 shows the comparisons of model selection criteria for the serum uric acid model. The comparisons were conducted only between 10-fold and n-fold cross-validation. The model selection criteria for two cross-validation methods looked similar to each other. However, n-fold cross-validation had higher R-square and all p-values of selected predictors were less than 0.2. Hence, the n-fold cross-validation method was used to construct the serum uric acid model.

There were many missing data for central obesity and waist percentile. The sample size of the model with central obesity and waist percentile was 54. The sample size of the model without central obesity and waist percentile was 75. Therefore, the model for 75 subjects without central obesity and waist percentile, and the subset analysis for 54 without central obesity and waist percentile, were constructed following on the previous procedures and logic of the α-klotho model. The tables of comparisons of model selection criteria for the three models were presented in Appendix A.

After comparisons of three serum uric acid models concerning about central obesity and waist percentile, based on a similar logic and analysis as α-klotho model, the best selected model for log (serum uric acid) used the n-fold cross-validation method and included central obesity and waist percentile. This model contained HbA1c, diabetic duration, age, ACR, Creatinine, Triglycerides, Cholesterol, HDL, central obesity, and waist percentile. The parameter estimates and p-values are shown in Table 19.

Table 19. Parameter estimates and p-values for the best selected log (serum uric acid) model

Parameter	Estimate	Standard Error	p-value
Intercept	1.902	0.608	0.003
HbA1c	-0.050	0.028	0.082
Diabetic duration	-0.015	0.012	0.198
Age	0.036	0.018	0.048
ACR	-0.002	0.002	0.200
Creatinine	0.652	0.268	0.020
TG	0.001	0.001	0.081
Cholesterol	-0.003	0.001	0.048
HDL	0.007	0.004	0.062
Central obesity	-2.915	1.324	0.033
Waist%	0.099	0.048	0.047

For the best selected log (serum uric acid) model, R-square = 0.447, AIC = -91.75, BIC = -138.40, Mallow’s Cp = 8.15, CV PRESS = 3.48.

3.6 REGRESSION DIAGNOSTICS

3.6.1 Collinearity

Collinearity was measured by Variance Inflation Factor (VIF). VIF=1 indicates perfect prediction and no effects of collinearity. Large deviations of VIF from 1 indicates collinearity. Generally, VIF ≥ 10 indicates the existence of collinearity issues.

3.6.1.1 Log (α-klotho) model

From Table 20, since all the VIF values were less than 10, no collinearity issue was presented for the log (α-klotho) model.

Table 20. VIF of log (α-klotho) model

	HbA1c	Growth velocity	TG	ACR	Cholesterol	HDL	Central obesity	Waist%
VIF	1.22	1.32	1.77	1.39	1.61	1.53	4.09	3.67

3.6.1.2 Log (eGFR) model

From Table 21, since all the VIF values were less than 10, no collinearity issue was presented for the log (α-klotho) model.

Table 21. VIF of log (eGFR) model

	HbA1c	Growth velocity	ACR	Cholesterol	DBP%	Creatinine	Central obesity
VIF	1.10	1.61	1.30	1.04	1.04	1.31	1.52

3.6.1.3 Log (serum uric acid) model

From Table 22, since all the VIF values were less than 10, no collinearity issue was presented for the log (serum uric acid) model.

Table 22. VIF of log (serum uric acid) model

	HbA1c	**Age**	**Cholesterol**	**TG**	**Creatinine**
VIF	1.39	1.95	1.56	1.75	1.60
	Central obesity	**HDL**	**Waist%**	**Diabetic Duration**	**ACR**
VIF	5.74	1.48	5.25	1.72	1.26

3.6.2 Linear regression assumptions checking

3.6.2.1 Linear regression assumptions

There are five assumptions for linear regression. They are:

1) Existence: For each specific combination of the fixed x's, y is a random variable with a certain probability distribution.
2) Independence: The y values are statistically independent of each other.
3) Linearity: The mean of y for each specific combination of x_1, x_2, x_3, x_4, … , x_k is a linear function of x_1, x_2, x_3, x_4, … , x_k.
4) Homoscedasticity: The variance of y is the same for any fixed combination of x_1, x_2, x_3, x_4, … , x_k.
5) Normality: For any fixed combination of x_1, x_2, x_3, x_4, … , x_k, the random variable y has a normal distribution.

The three models should be checked whether they met all linear regression assumptions.

3.6.2.2 Existence and Independence

From the study design, children were recruited separately and no children had blood relationships. We assumed each of three dependent variables had a certain probablity distribution. Therefore, existence and independence assumptions were met for all three models.

3.6.2.3 Linearity and Homoscedasticity

Linearity and homoscedasticity could be assessed by using a plot of residuals versus predicted values. The residuals should be small and symmetric around 0. In addition, no obvious patterns of residuals suggested the linearity.

Figure 2 displays the scatter plot of residuals by predicted values for the log (α-klotho) model. Almost all residuals were between -0.5 and 0.5. They were randomly distributed around 0 and no other pattern was obviously presented. Hence, linearity and homoscedasticity assumptions were met for the log (α-klotho) model.

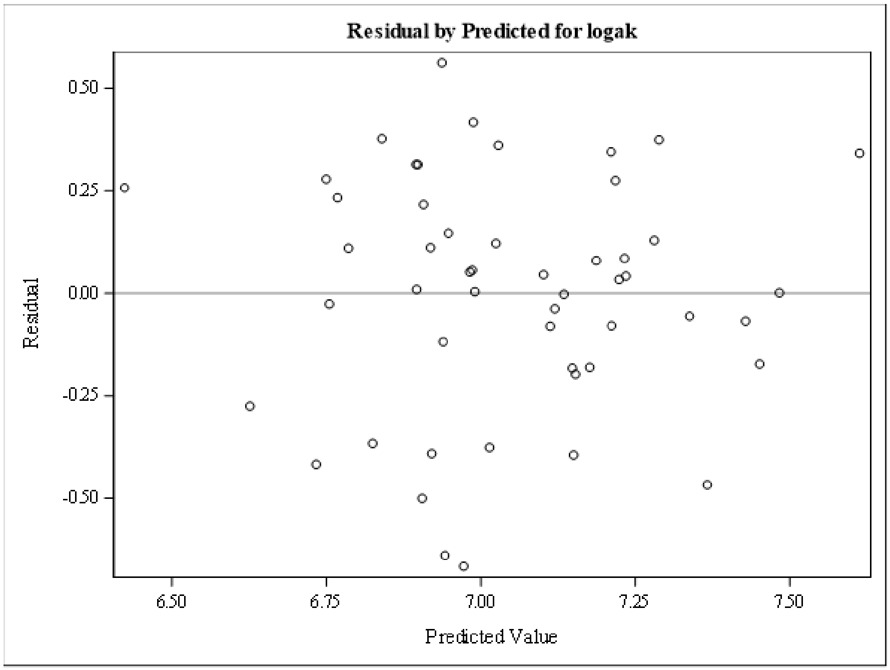


Figure 2. Scatter plot of residuals by predicted values for log (α-klotho) model

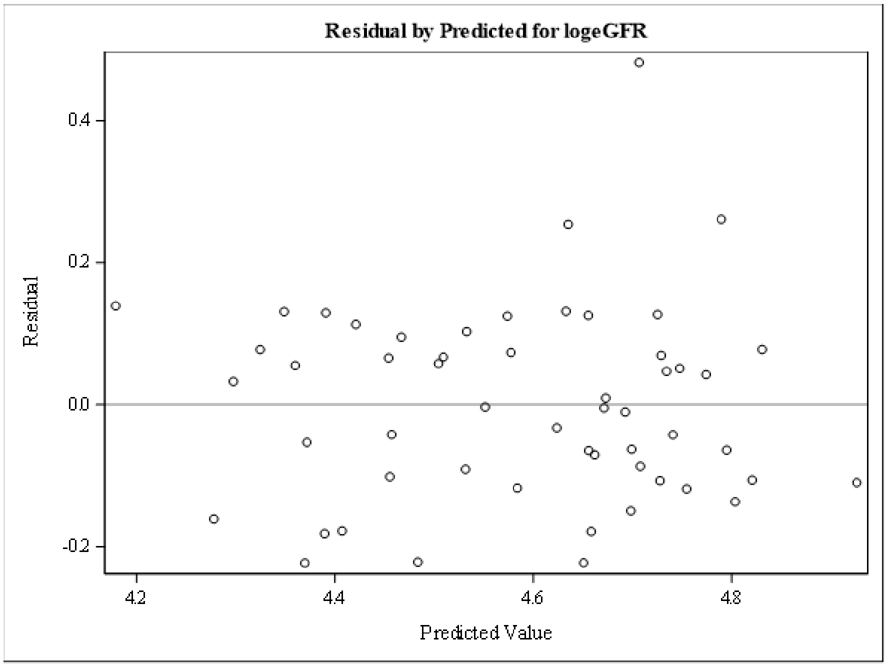


Figure 3. Scatter plot of residuals by predicted values for log (eGFR) model

Figure 3 shows the scatter plot of residuals by predicted values for the log (eGFR) model. Most residuals fell between -0.2 to 0.2 and there was no obvious pattern for the residuals. Thus, the linearity assumption was met for the log (eGFR) model.

Figure 4 is the scatter plot of residuals by predicted values for the log (serum uric acid) model. Most residuals were between -0.4 to 0.4 and they were symmetric and randomly distributed around 0. Also, there was no obvious pattern for the residuals. Hence, linearity and homoscedasticity assumptions were met for the log (serum uric acid) model.

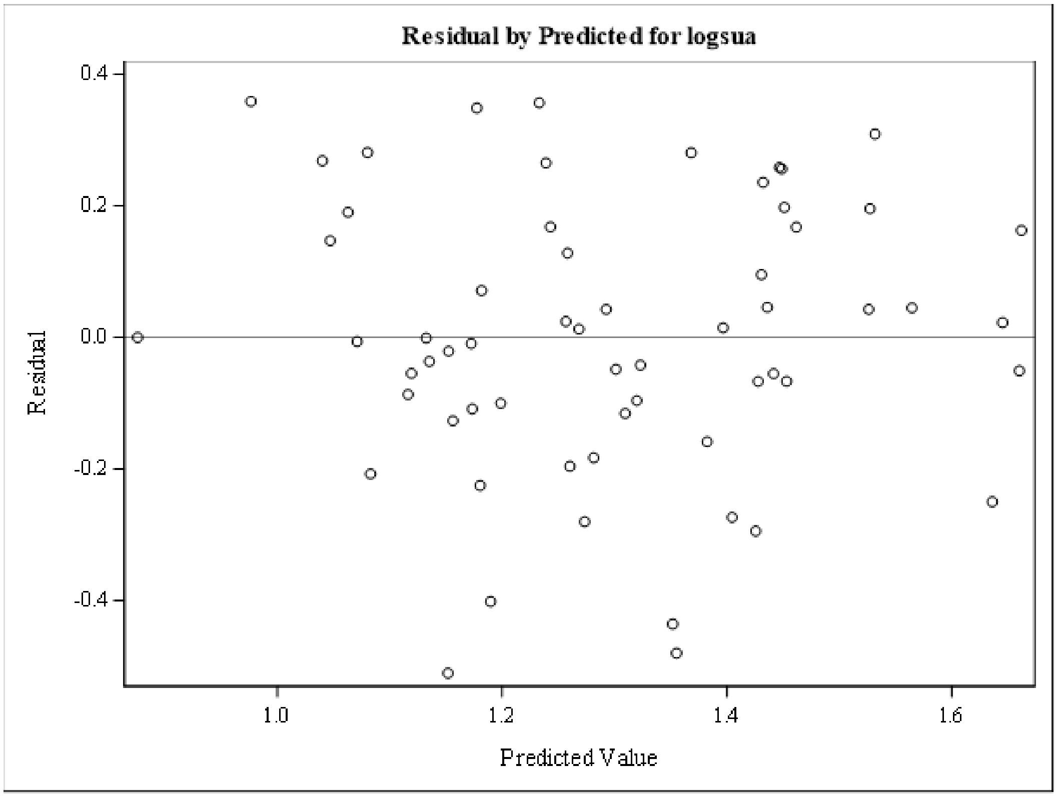


Figure 4. Scatter plot of residuals by predicted values for log (serum uric acid) model

3.6.2.4 Normality

The quantile-quantile plot (QQ plot) was used to access the normality assumptions for the three models.

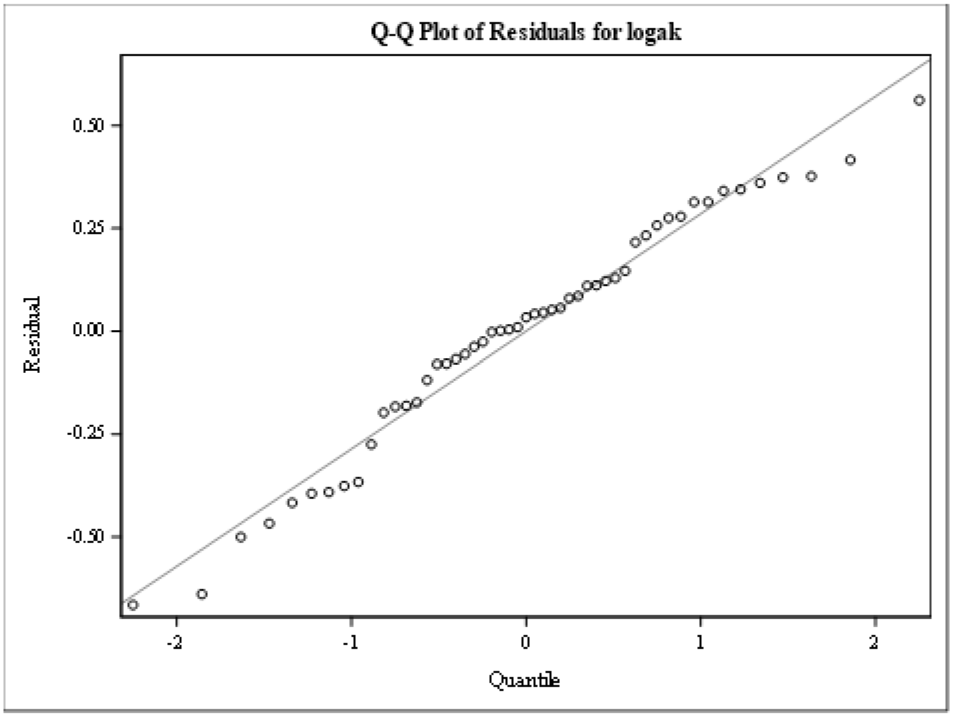


Figure 5. QQ plot of residuals for log (α-klotho) model

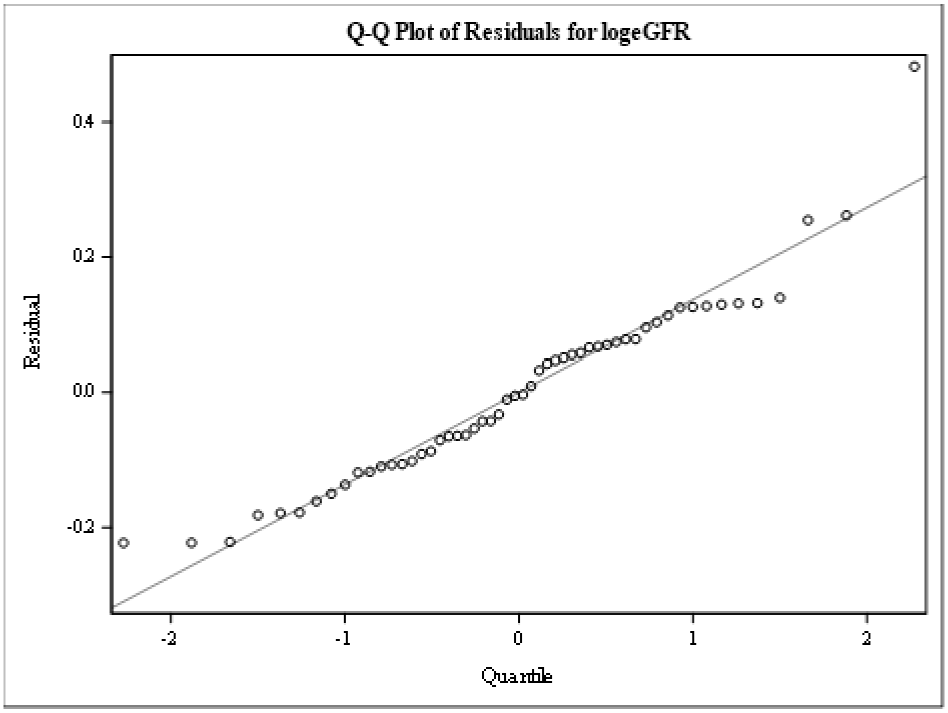


Figure 6. QQ plot of residuals for log (eGFR) model

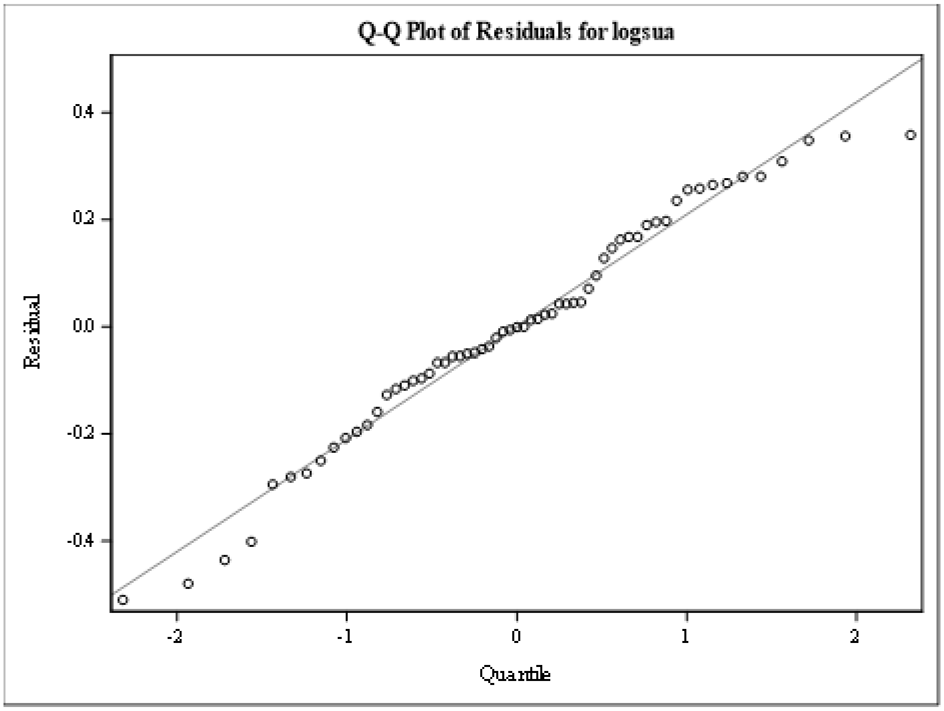


Figure 7. QQ plot of residuals for log (serum uric acid) model

From Figure 5-7, normality assumptions were met for all three models.

In all, the log (α-klotho) model, log (eGFR) model and log (serum uric acid) model all met the five assumptions of linear regression.

3.6.3 Problematic points

Different types of plots and methods could be used to detect problematic points. Problematic points included high leverage points, outliers, and influential points. High leverage points could be examined by a plot of leverage analysis. Outliers could be examined by a plot of Studentized Residuals. Influential points could be examined by a plot of Cook's Distance versus observation or DFFITS versus observation. The cut point formulas and values of three methods for the three models are shown in Table 23.

Table 23. Cut point formulas and values for the three models

Leverage	Leverage	Outlier	Cook's D	|DFFITS|
Cut point formula	2(k+1)/n	±2 of Studentized Residual	4/n	2 √ [(k+1)/n]
Values for α-klotho model	0.353	±2	0.078	0.840
Values for eGFR model	0.296	±2	0.074	0.770
Values for serum uric acid model	0.361	±2	0.066	0.849

k is number of predictors, n is number of observations in the model.

3.6.3.1 Log (α-klotho) model

Figure 8 displays plots of problematic points for the log (α-klotho) model. There were two outliers, subject 16 and subject 69, whose studentized residuals were larger than the cut point ±2. There was one high leverage point, subject 97, whose leverage was larger than cut point 0.353. There were three subjects (subject 16,69 and 91), who had high Cook's D values above the cut point 0.078. In addition, subject 16, 69, 79 and 91 had the DEFFITS values below the cut point -0.840. Generally, there were few problematic points. Thus, the log (α-klotho) model was considered as a good predictive model. Fit diagnostics for other methods and residuals by each predictor plots are shown in Appendix A.

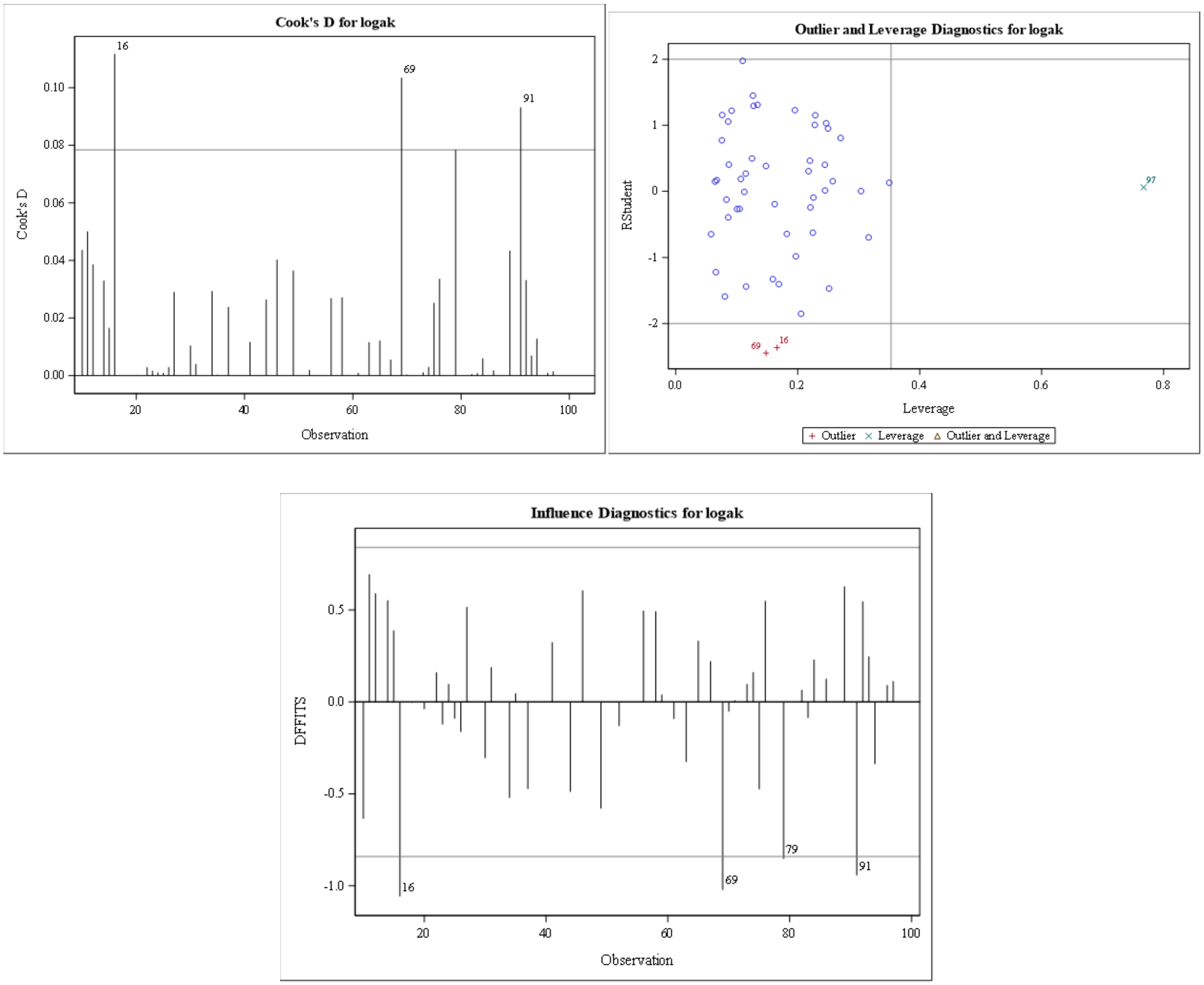


Figure 8. Plots of problematic points for log (α-klotho) model

3.6.3.2 Log (eGFR) model

Figure 9 displays plots of problematic points for the log (eGFR) model. There was one outlier, subject 35, whose studentized residual was larger than the cut point ±2. There were two high leverage points, subject 47 and 75, whose leverages were larger than cut point 0.296. There were two subjects (subject 35 and 92), who had high Cook’s D values above the cut point 0.074. In addition, subject 35 and subject 92 had the DEFFITS values below or above the cut point ±0.770. Generally, there were few problematic points. Hence, the log (eGFR) model was considered as a good predictive model. Fit diagnostics for other methods and residuals by each predictor plots are shown in Appendix A.

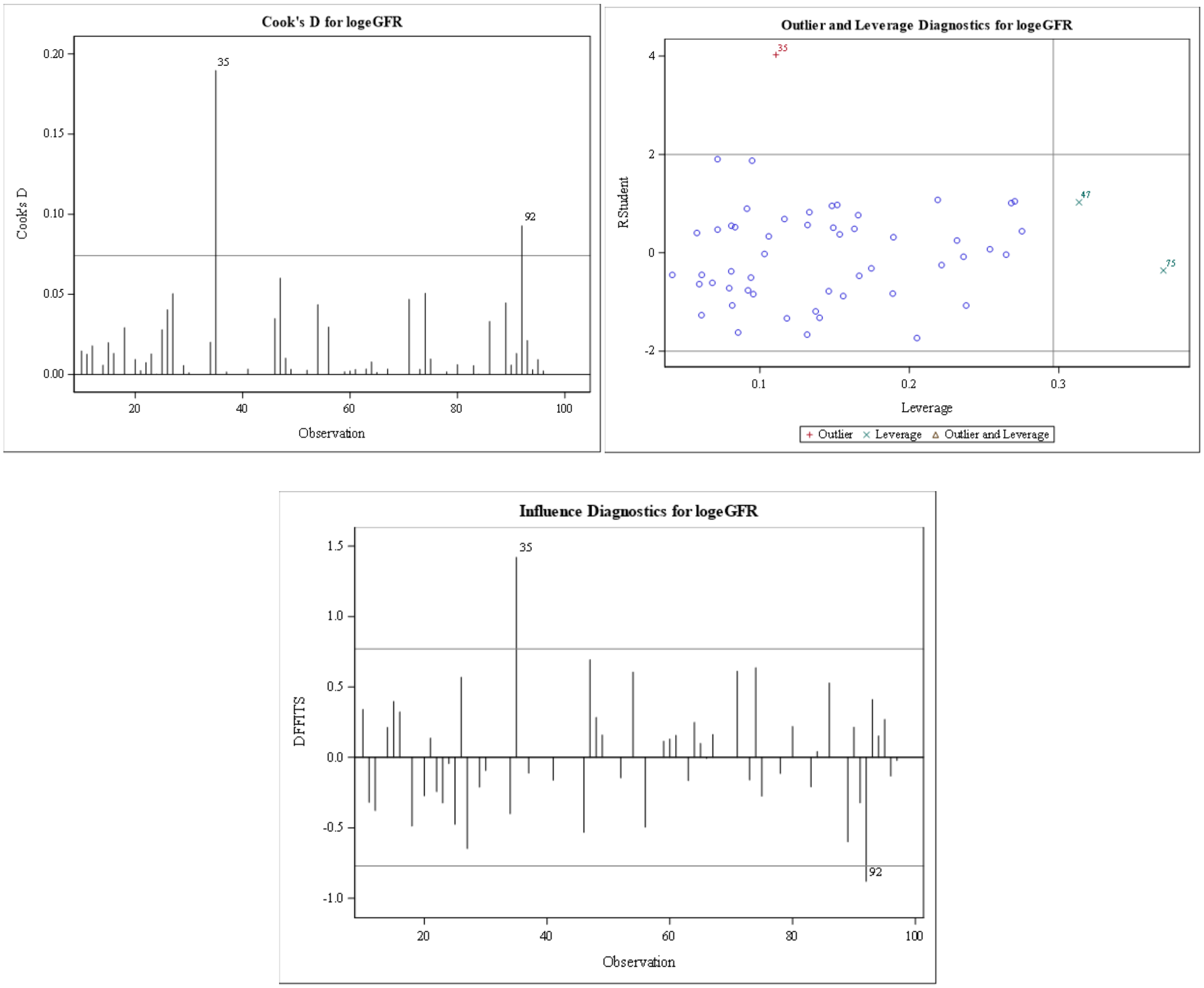


Figure 9. Plots of problematic points for log (eGFR) model

3.6.3.3 Log (serum uric acid) model

Figure 10 displays plots of problematic points for the log (serum uric acid) model. There were three outliers, subject 11, 41, and 93, whose studentized residuals were larger than the cut point ±2. There were two high leverage points, subject 47 and 97, whose leverages were larger than cut point 0.361. There were three subjects (subject 52, 75, and 93) who had high Cook's D values above the cut point 0.066. In addition, subject 52, 75, and 93 had DEFFITS values below or above the cut point ±0.849. Generally, there were few problematic points, so the log (serum uric acid) model was considered as a good predictive model. Fit diagnostics for other methods and residuals by each predictor plots are shown in Appendix A.

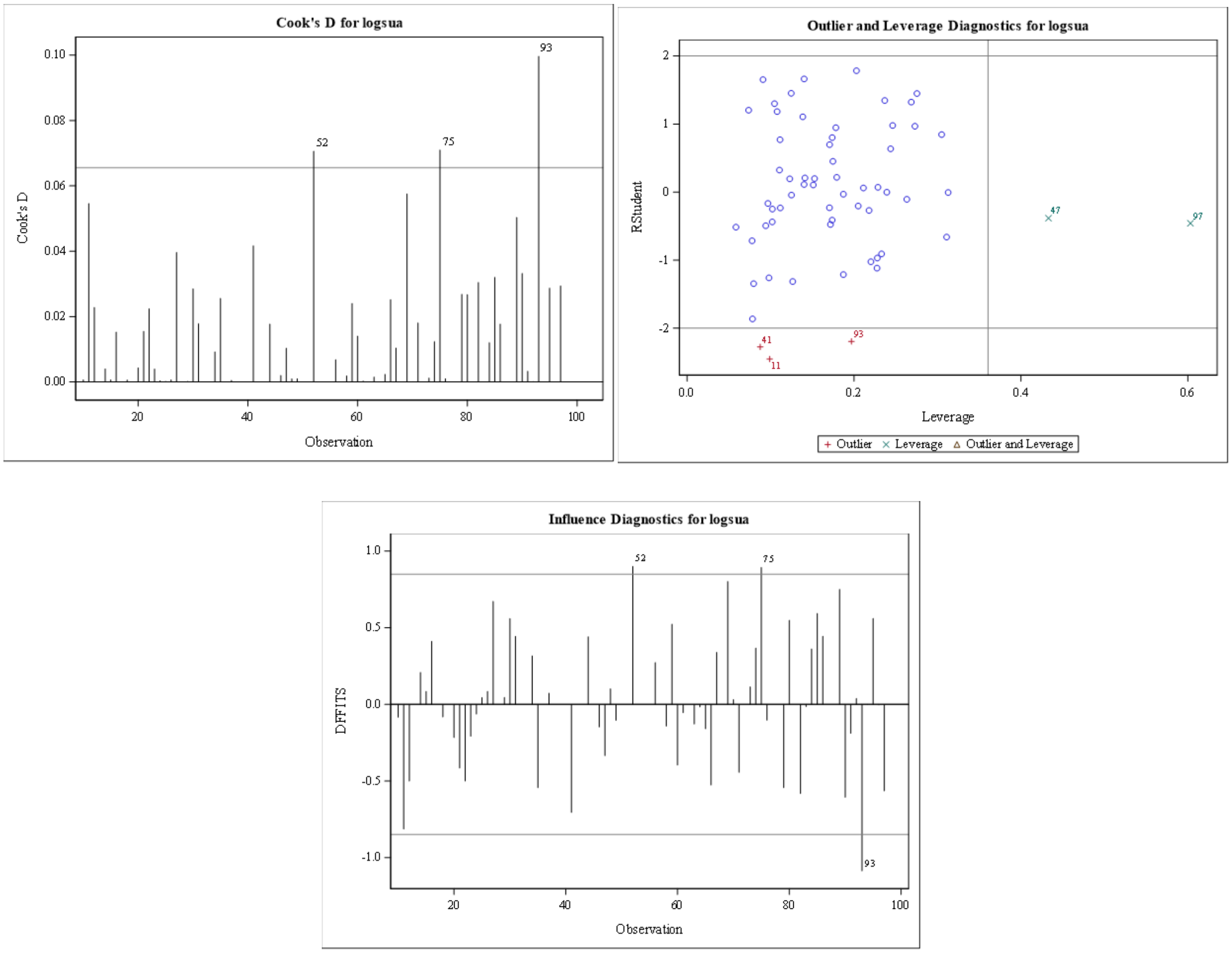


Figure 10. Plots of problematic points for log (serum uric acid) model

Table 24. Comparisons of model selection criteria for three models

	R-square	AIC	BIC	Mallow’s Cp	CV PRESS
Log (α-klotho) model	0.4236	-51.03	-88.55	4.33	5.11
Log (eGFR) model	0.6176	-140.23	-185.62	0.18	1.24
Log (serum uric acid) model	0.4467	-91.75	-138.40	8.15	3.48

3.7 ACR MODELS

ACR reflects the albuminuria level, which is the gold standard for predicting and screening the progression of DN. Hence, it is important to predict the ACR level using other variables and potential risk biomarkers.

3.7.1 Transformation of ACR

Because ACR was treated as the dependent variable in the ACR model, distribution needed to be checked to assess whether a log transformation was necessary.

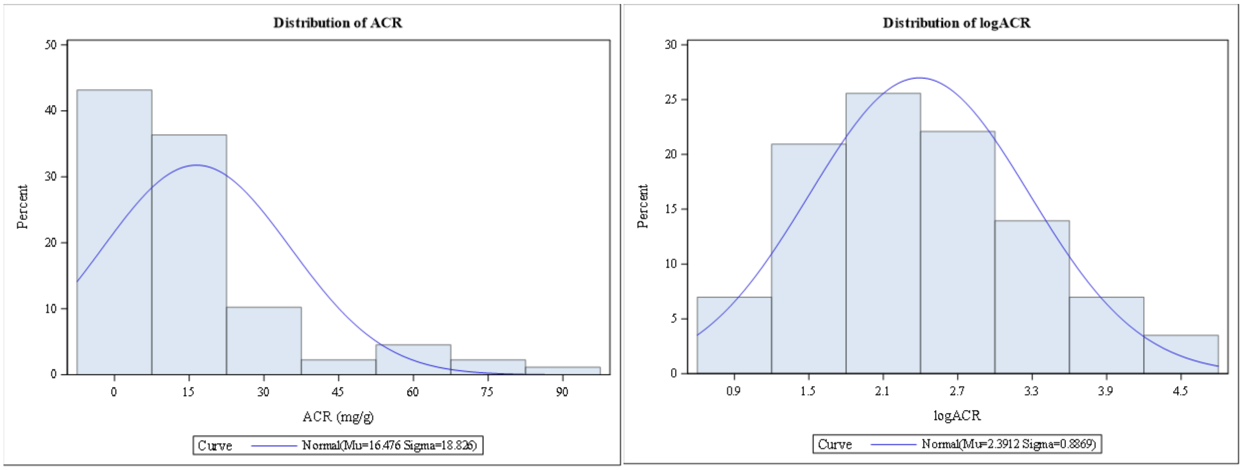


Figure 11. Distribution of ACR and the log transformation

Figure 11 shows histograms of the distribution of ACR and the log transformation. The left picture is the original ACR distribution and the right picture is the ACR distribution with log transformation. Since the original ACR distribution was extremely positive-skewed and log (ACR) distribution seemed normally-distributed, the log transformation of ACR was used as the dependent variable in the following modelling steps.

3.7.2 ACR model with potential biomarkers

After determining the appropriate transformation of the dependent variable, multiple linear regression models were constructed to see which variables were considered important predictors for ACR.

Recent studies showed that ACR was related to age, diabetic duration years, and HbA1c. As a result, those three variables were forced into the ACR model. Since we wanted to see whether the three potential risk biomarkers were important predictors for the ACR level, those three biomarkers along with three forced variables were treated as the independent variables in the first ACR model. BMI percentile, waist circumference percentile and central obesity were excluded from the model because they had a substantial number of missing data and the sample size would increase by excluding them.

Based on the previous results and discussions, the n-fold cross-validation method and backwards elimination were used to construct and select the ACR model.

Table 25. Parameter estimates and p-values for the log(ACR) model with three biomarkers

Parameter	Estimate	Standard Error	p-value
Intercept	1.384	1.005	0.173
HbA1c	0.191	0.091	0.040
Diabetic Duration	-0.009	0.032	0.788
Age	-0.026	0.046	0.578

Table 25 shows parameter estimates and p-values for this ACR model. For this model, R-square = 0.0751, AIC = 62.56, BIC = -2.77, Mallow’s Cp = 2.83, CV PRESS = 60.68.

The results showed that none of the three risk biomarkers (α-klotho, eGFR, serum uric acid) were related to ACR and they were not important predictors for ACR.

From parameter estimates and model selection criteria, this model could not be considered a good predictive model. The R-square was at an extremely low level, which indicated that only 7.5% variability of log (ACR) could be explained by selected variables. CV PRESS was extremely high, which indicated that this model had high residuals and was not predictive for new observations. Thus, we may seek other ACR models which can predict ACR better and involve other relevant predictors.

3.7.3 ACR model with biomarkers and other variables

Some other variables, such as SBP percentile, DBP percentile, LDL, HDL and Triglycerides, were considered candidate variables for predicting ACR. Hence, these variables, along with three forced variables and three potential risk biomarkers, were included as candidate variables in the ACR model to see whether those variables and biomarkers were important predictors for ACR.

Based on the previous results and discussions, the n-fold cross-validation method and backwards elimination were used to construct and select the ACR model.

Table 26. Parameter estimates and p-values for the log(ACR) model with biomarkers and other variables

Parameter	Estimate	Standard Error	p-value
Intercept	-0.856	1.337	0.525
HbA1c	0.227	0.094	0.019
Diabetic Duration	-0.021	0.033	0.524
Age	-0.0002	0.050	0.996
SBP%	0.016	0.007	0.021
DBP%	-0.021	0.010	0.035
TG	0.004	0.002	0.029
HDL	0.032	0.011	0.008

Table 26 shows parameter estimates and p-values for the ACR model with three biomarkers and other variables. For this model, R-square = 0.2751, AIC = 55.35, BIC = -2.20, Mallow's Cp = 4.93, CV PRESS = 55.23.

The results showed that none of the three risk biomarkers (α-klotho, eGFR, serum uric acid) were related to ACR. However, SBP percentile, DBP percentile, Triglycerides, and HDL were considered important predictors for ACR.

For the ACR model with biomarkers and other variables, there was no collinearity issue. All linear regression assumptions were met. However, there were several problematic points for this model, which indicated this model was not as predictive as three biomarkers' models. Relevant graphs are displayed in the Appendix A.

4.0 DISCUSSION AND CONCLUSION

For three potential biomarkers, separate models were constructed using demographic and clinical data. Log transformation of the biomarkers were made and normality was achieved.

α-klotho was negatively correlated with diabetic duration years and HbA1c, which indicated that α-klotho was associated with glycemic control and it could be a potential early risk biomarker for diabetic complications in children with T1D. However, α-klotho was not significantly correlated with ACR in this study. Future studies including longitudinal follow-up of subjects need to be done to assess the relationship between ACR and α-klotho. For the multiple regressions of log (α-klotho) model, after model selection and cross-validation, the final model included HbA1c, growth velocity, Triglycerides, total Cholesterol, HDL and central obesity. HbA1c was still an important predictor in the multiple regression model and it reflected that glycemic control plays a vital role in predicting α-klotho. Some clinical predictors (Triglycerides, total Cholesterol, HDL) that were not significant predictors in univariable regression analysis did become significant predictors in the final multiple regression models. It indicated that univariable regression analysis alone was not sufficient and predictive. Multiple regression analysis was necessary and may predict the model better than univariate. From the selected log (α-klotho) model, R-square was largely improved (R^2=0.37), which meant this model had a relative better fit compared to individual univariable regression models.

eGFR was not correlated with HbA1c and ACR, which indicated that eGFR was not strongly associated with glycemic control and it may not be a potential early risk biomarker for diabetic complications in youth with T1D. However, the p-value between eGFR and HbA1c was close to 0.05, which indicated the weak association. Future studies about longitudinal follow-ups of subjects needed to be done to assess the relationship between eGFR and HbA1c. After fitting the multiple regression model for log (eGFR) and conducting model selections, HbA1c and ACR were both included in the selected model. Hence, glycemic control and albuminuria were important factors in predicting eGFR. Also, eGFR may be a potential early risk biomarker for DN and diabetic complications in children with T1D.

Serum uric acid was negatively correlated with HbA1c. But the p-values were both between 0.05 and 0.1 and they were not high p-values. Significant correlations between eGFR and HbA1c or, ACR could be possibly derived if more subjects were enrolled in the study. After fitting the multiple regression model for log (serum uric acid) and conducting model selections, HbA1c and ACR were both included in the selected model. Hence, glycemic control and albuminuria were important factors in predicting serum uric acid.

The three final models all contained HbA1c and central obesity. HbA1c was correlated with three novel biomarkers. HbA1c measures the level of glycemic control and existence of HbA1c in the final three models indicate the close relationships between glycemic control and potential biomarkers. Central obesity was measured by waist/height ratio and it was highly related to the three novel biomarkers. Moreover, serum-related variables such as cholesterol, Creatinine, HDL and Triglycerides were considered important in predicting the three biomarkers. ACR, which measures the albuminuria, the current gold standard for predicting the progression of diabetic nephropathy, was an important predictor for eGFR and serum uric acid models.

Table 24 shows the results for the different model selection criteria of the three models. Log (eGFR) model is the most predictive model among the three because it had the highest R-square, which was larger than 0.5. Also, the log (eGFR) model had the lowest AIC, BIC, Mallow's Cp, and CV PRESS, which demonstrated that log (eGFR) model was relatively a precise model of predicting the dependent variable and estimating parameters. For the other two models, log (α-klotho) model had the lowest R-square, highest AIC, BIC, and CV PRESS among the three models, which indicated that α-klotho model was the least predictive model.

For all three models, the R-square was not that large. The largest R-square was 0.62, which indicated that there was still 38% of variance could not explained by selected variables. In addition, the Mallow's Cp of three models were all smaller than the number of predictors plus 1, which indicate a tendency of overfitting. By looking at the database, the sample size of all three models were less than 70 and there were more than 15 variables in the maximum model. Thus, overfitting issues could exist in the selection of the final models. In the future studies, more children should be recruited to increase the sample size and avoid overfitting issues. More clinical predictors could be measured and added to the study to derive more predictive models with higher R-square. Longitudinal analysis with follow-up data are needed to test the time trend and whether other time-dependent models are more appropriate, such as linear mixed model and GEE.

For the ACR model with three potential risk biomarkers and other variables, after model selection, none of three biomarkers (α-klotho, eGFR and serum uric acid) were considered important predictors of ACR. SBP percentile, DBP percentile, HDL and Triglycerides were important predictors for predicting ACR. However, the R-square of the model was relatively low (0.275) and CV PRESS was larger than 50, which indicated that this ACR model was not an accurate model and was not predictive for new observations. Due to the small sample size and restricted numbers of variables, this ACR model may not reflect the true relationship between ACR and three potential biomarkers. In the future study, more children should be recruited to increase the sample size. In addition, more variables with fewer missing data need to be considered in order to elevate the R-square. Moreover, follow-up data should be collected and longitudinal analysis need to be performed to assess the relationship between ACR and potential biomarkers.

APPENDIX A: SUPPLEMENTAL RESULTS

Table 27. Characteristics of children with diabetic nephropathy by ACR normal and abnormal groups

Variable	**ACR normal group (ACR < 30 mg/g) (n = 74)**	**ACR abnormal group (ACR ≥ 30 mg/g) (n=15)**	**p-value**
Sex (Male)		5 (33.3)	0.10
Male	42 (56.8)	5 (33.3)	
Female	32 (43.2)		
Race (White)	66 (89.2)	13 (86.7)	0.63
Age (years)	16.1 ± 3.0	15.3 ± 2.7	0.35
Diabetes duration (years)	7.1 ± 3.6	7.6 ± 4.8	0.65
Average SBP (mm Hg)	114.3 ± 8.3	111.5 ± 7.4	0.23
Average DBP (mm Hg)	72.5 ± 5.1	70.1 ± 7.1	0.33
BMI (kg/m^2)	23.7 ± 4.8	22.7 ± 6.1	0.49
Waist circumference (cm)	81.0 ± 12.0	72.0 ± 12.5	0.02
Waist/height ratio	0.49 ± 0.06	0.45 ± 0.08	0.07
Growth velocity (cm/year)	2.6 ± 2.7	3.6 ± 2.9	0.20
Hemoglobin A1c (%)	8.0 ± 1.3	8.5 ± 1.4	0.15
Insulin dose (u/kg/d)	0.88 ± 0.26	0.93 ± 0.26	0.66
α-klotho (pg/mL)	1270.4 ± 540.5	1387.6 ± 648.3	0.37
Cystatin C (mg/L)	0.83 ± 0.14	0.77 ± 0.12	0.16
Estimated GFR (ml/kg/1.73m^2)	97.56 ± 19.81	105.50 ± 19.31	0.16
Serum Uric Acid (mg/dL)	3.83 ± 0.98	3.57 ± 0.92	0.34
Creatinine (mg/dL)	1.95 ± 10.76	0.69 ± 0.15	0.15
Triglycerides (mg/dL)	98.2 ± 61.33	126.0 ± 109.4	0.34
Total cholesterol (mg/dL)	164.9 ± 26.89	176.7 ± 30.5	0.13
LDL (mg/dL)	90.0 ± 24.67	96.6 ± 20.7	0.34
HDL (mg/dL)	56.3 ±	60.9 ± 9.7	0.15

Mann-Whitney U test or Students' t-test for continuous variables and Fisher's exact test or chi-square test for categorical variables.

Table 28. Characteristics of children with diabetic nephropathy by ACR quartiles

Variable	1st ACR quartile (n=21)	2nd ACR quartile (n=23)	3rd ACR quartile (n=23)	4th ACR quartile (n=22)
Sex (Male)	14 (66.6)	10 (43.5)	12 (52.2)	11 (50.0)
Race (White)	20 (95.2)	20 (87.0)	21 (91.3)	18 (81.8)
Age (years)	16.2 ± 2.0	15.7 ± 3.2	16.3 ± 3.1	15.7 ± 3.2
Diabetes duration (years)	7.2 ± 3.6	7.0 ± 3.5	7.1 ± 4.2	7.4 ± 4.1
Average SBP (mm Hg)	111.0 ± 7.7	115.0 ± 8.6	116.5 ± 8.7	112.5 ± 7.1
Average DBP (mm Hg)	70.7 ± 5.2	73.8 ± 4.5	72.8 ± 5.5	71.4 ± 6.4
Height (cm)	167.7 ± 13.6	163.6 ± 10.9	161.8 ± 11.1	162.0 ± 14.1
Weight (kg)	67.6 ± 19.1	66.4 ± 19.8	62.5 ± 18.4	57.6 ± 18.3
BMI (kg/m^2)	24.0 ± 4.3	24.3 ± 5.2	23.4 ± 4.6	22.4 ± 5.9
Waist circumference (cm)[a]	81.9 ± 10.3	83.2 ± 12.3	78.3 ± 12.4	74.6 ± 14.1
WHR	0.48 ± 0.05	0.51 ± 0.07	0.48 ± 0.05	0.46 ± 0.08
Growth velocity (cm/year)	2.1 ± 2.1	2.8 ± 3.3	3.0 ± 2.7	3.3 ± 2.8
Hemoglobin A1c (%)	7.8 ± 1.3	8.0 ± 1.3	8.1 ± 1.3	8.3 ± 1.4
Insulin dose (u/kg/d)	0.87 ± 0.30	0.97 ± 0.28	0.85 ± 0.22	0.87 ± 0.23
α-klotho (pg/mL)	1197.1 ± 315.3	1380.1 ± 651.1	1292.7 ± 637.6	1293.0 ± 584.6
Cystatin C (mg/L)[a]	0.89 ± 0.13	0.80 ± 0.15	0.80 ± 0.15	0.79 ± 0.11
Estimated GFR (ml/kg/1.73m^2)	89.47 ± 14.33	103.43 ± 25.16	100.87 ± 18.97	102.85 ± 18.23
Serum Uric Acid (mg/dL)	3.88 ± 1.09	3.79 ± 0.77	3.84 ± 0.90	3.65 ± 1.14
Creatinine (mg/dL)	0.70 ± 0.15	0.68 ± 0.15	4.60 ± 18.90	0.69 ± 0.17
Triglycerides (mg/dL)	82.6 ± 43.0	119.7 ± 87.1	98.4 ± 46.3	109.2 ± 93.6
Total cholesterol (mg/dL)	166.5 ± 19.7	169.6 ± 26.2	159.8 ± 34.1	172.0 ± 28.8
LDL (mg/dL)	89.1 ± 27.1	97.3 ± 21.6	85.0 ± 27.0	93.1 ± 19.5
HDL (mg/dL)	56.8 ± 10.2	53.6 ± 10.2	57.1 ± 13.4	60.9 ± 10.9

[a] Significant linear trend, $p < 0.05$

Table 29. Characteristics of children with diabetic nephropathy by low HbA1c and high HbA1c groups

Variable	Low HbA1c group (HbA1c <7.5%) (n=32)	High HbA1c group (HbA1c ≥7.5%) (n=66)	p-value
Sex (Male)	18 (56.3)	32 (48.5)	0.47
Race (White)	30 (96.8)	57 (87.7)	0.26
Age (years)	16.6 ± 3.2	15.4 ± 2.7	0.04
Diabetes duration (years)	6.3 ± 4.4	7.4 ± 3.5	0.06
Average SBP (mm Hg)	114.9 ± 7.3	112.8 ± 8.9	0.26
Average DBP (mm Hg)	72.7 ± 4.5	71.6 ± 6.0	0.40
Height (cm)	166.8 ± 11.3	161.4 ± 12.8	0.04
Weight (kg)	64.4 ± 15.0	62.0 ± 20.5	0.29
BMI (kg/m^2)	23.1 ± 4.1	23.5 ± 5.3	0.74
Waist circumference (cm)	79.9 ± 9.1	78.5 ± 13.6	0.58
WHR	0.48 ± 0.05	0.48 ± 0.07	0.98
Growth velocity (cm/year)	2.5 ± 2.5	3.0 ± 2.8	0.58
ACR (mg/g)	12.7 ± 14.4	27.2 ± 70.4	0.20
Insulin dose (u/kg/d)	0.81 ± 0.25	0.94 ± 0.26	0.03
Soluble α-klotho (pg/mL)	1485.8 ± 628.3	1216.7 ± 517.0	0.03
Cystatin C (mg/L)	0.85 ± 0.16	0.79 ± 0.12	0.05
Estimated GFR (ml/kg/1.73m^2)	94.11 ± 18.44	103.02 ± 21.29	0.07
Serum Uric Acid (mg/dL)	4.28 ± 0.99	3.55 ± 0.83	0.0003
Creatinine (mg/dL)	3.64 ± 16.27	0.66 ± 0.13	0.04
Triglycerides (mg/dL)	88.6 ± 47.0	109.5 ± 77.6	0.16
Total cholesterol (mg/dL)	162.5 ± 21.1	167.7 ± 29.8	0.32
LDL (mg/dL)	90.6 ± 19.4	90.1 ± 25.8	0.93
HDL (mg/dL)	54.8 ± 11.3	58.1 ± 11.7	0.19

Mann-Whitney U test or Students' t-test for continuous variables and Fisher's exact test or chi-square test for categorical variables.

Table 30. Correlations between Hemoglobin A1c and physical characteristics, glycemic control, urine-related, serum related variables adjusting for age (Spearman correlation)

Variables	Adjusted r	p-value
Diabetes duration (years)	0.26	0.01
Average SBP (mm Hg)	-0.02	0.83
Average DBP (mm Hg)	0.11	0.28
Height (cm)	-0.06	0.58
Weight (kg)	-0.04	0.69
BMI (kg/m^2)	0.02	0.86
Waist circumference (cm)	0.03	0.80
WHR	0.07	0.54
Growth velocity (cm/year)	-0.11	0.30
ACR (mg/g)	0.16	0.15
Insulin dose (u/kg/d)	0.25	0.03
Soluble α-klotho (pg/mL)	-0.32	0.004
Serum Uric Acid (mg/dL)	-0.22	0.04
Cystatin C (mg/L)	-0.20	0.06
Estimated GFR (ml/kg/1.73m^2)	0.21	0.06
Creatinine (mg/dL)	-0.21	0.05
Triglycerides (mg/dL)	0.31	0.002
Total cholesterol (mg/dL)	0.09	0.38
LDL (mg/dL)	0.004	0.97
HDL (mg/dL)	0.09	0.38

Table 31. Correlations between ACR and physical characteristics, glycemic control, urine-related, serum related variables adjusting for age (Spearman correlation)

Variables	Adjusted r	p-value
Diabetes duration (years)	-0.004	0.97
Average SBP (mm Hg)	0.17	0.10
Average DBP (mm Hg)	0.07	0.50
Height (cm)	-0.13	0.24
Weight (kg)	-0.18	0.09
BMI (kg/m^2)	-0.17	0.11
Waist circumference (cm)	-0.28	0.02
WHR	-0.27	0.03
Growth velocity (cm/year)	0.17	0.12
Hemoglobin A1c (%)	0.16	0.15
Insulin dose (u/kg/d)	-0.03	0.80
Soluble α-klotho (pg/mL)	0.02	0.84
Serum Uric Acid (mg/dL)	-0.04	0.74
Cystatin C (mg/L)	-0.21	0.06
Estimated GFR (ml/kg/1.73m^2)	0.21	0.06
Creatinine (mg/dL)	0.04	0.69
Triglycerides (mg/dL)	0.12	0.26
Total cholesterol (mg/dL)	0.05	0.65
LDL (mg/dL)	-0.03	0.80
HDL (mg/dL)	0.16	0.14

Table 32. Correlations between estimated GFR and physical characteristics, other potential biomarkers adjusting for age (Spearman correlation)

Variables	Adjusted r	p-value
Diabetes duration (years)	-0.17	0.13
Average SBP (mm Hg)	0.10	0.37
Average DBP (mm Hg)	0.23	0.03
Height (cm)	-0.37	0.0005
Weight (kg)	-0.11	0.30
BMI (kg/m^2)	0.10	0.34
Waist circumference (cm)	0.11	0.40
WHR	0.33	0.007
Growth velocity (cm/year)	-0.32	0.003
Soluble α-klotho (pg/mL)	-0.19	0.11
Serum Uric Acid (mg/dL)	-0.36	0.0007

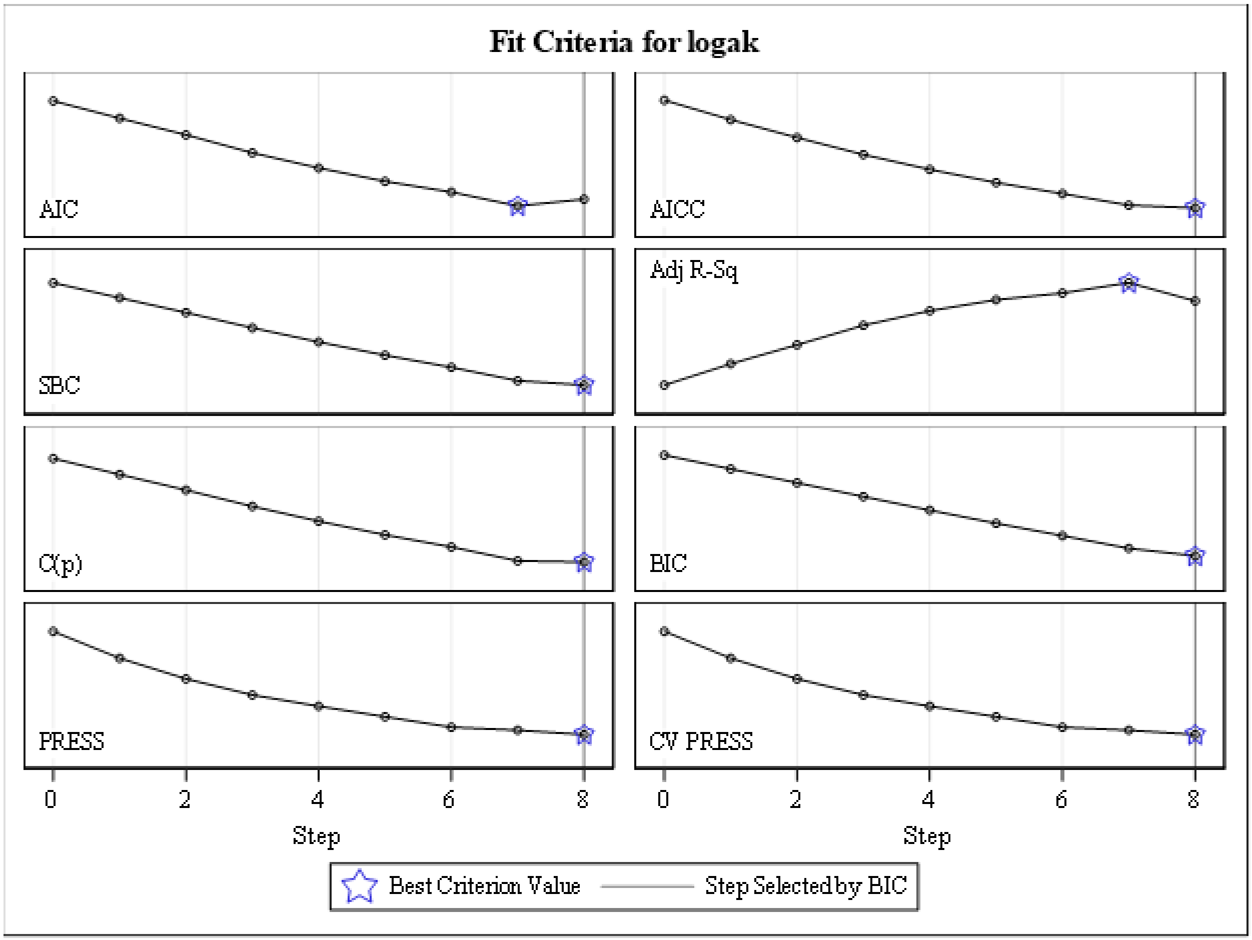


Figure 12. Fit criteria for the best selected log (α-klotho) model

Table 33. Model selection criteria with or without central obesity and waist% for eGFR model

	N	R^2	Adjusted R^2	AIC	BIC	Mallow's Cp	CV PRESS
With central obesity and waist%	52	0.6176	0.5568	-140.23	-185.62	0.18	1.24
Without central obesity and waist%	70	0.2700	0.2005	-165.25	-232.98	4.03	2.35

Table 34. Model selection criteria without central obesity and waist% for non-missing dataset for eGFR model

	N	R^2	Adjusted R^2	AIC	BIC	Mallow's Cp	CV PRESS
Without central obesity and waist% for non-missing dataset	52	0.5744	0.5176	-136.66	-183.83	0.14	1.32

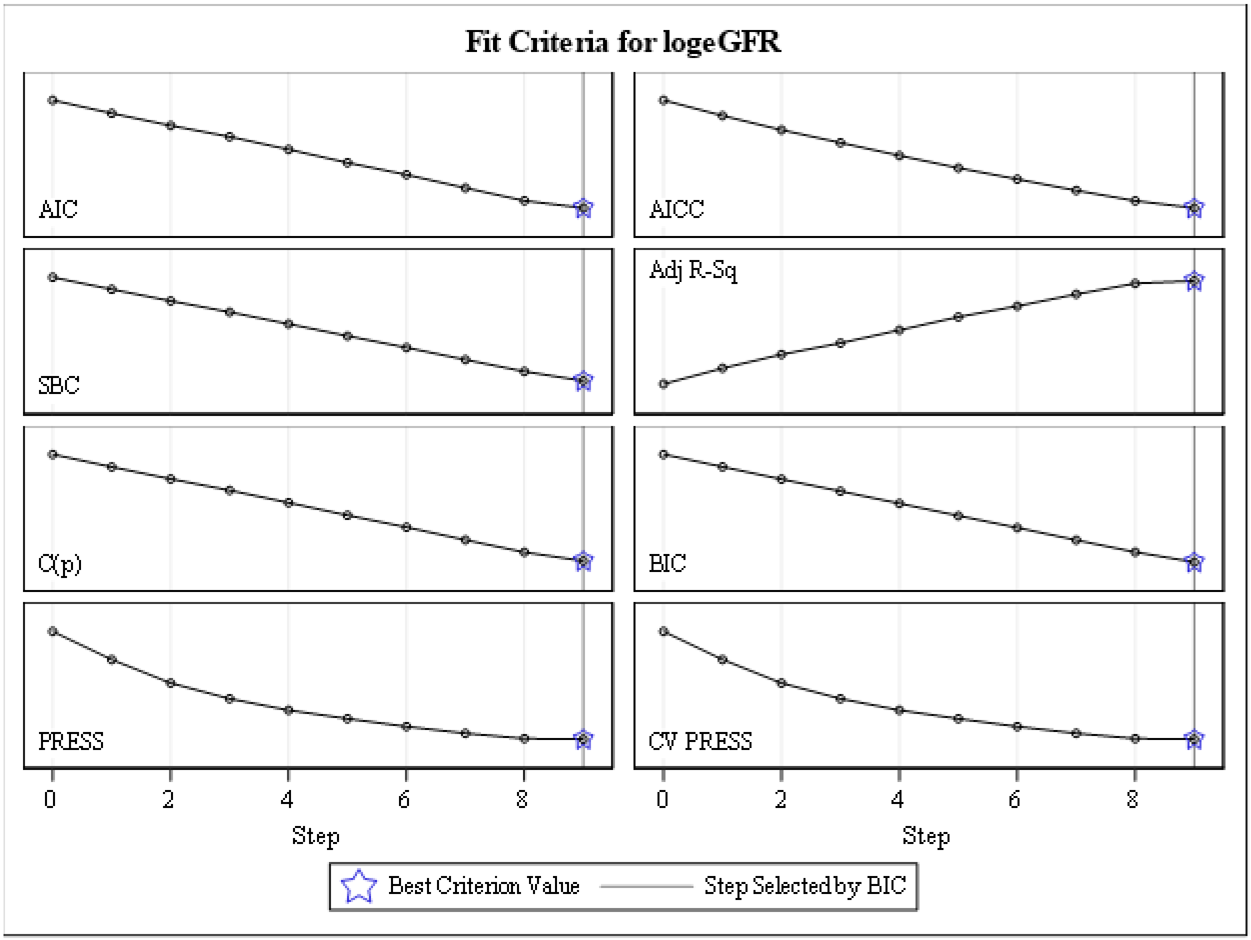


Figure 13. Fit criteria for the best selected log (eGFR) model

Table 35. Model selection criteria with or without central obesity and waist% for serum uric acid model

	N	R^2	Adjusted R^2	AIC	BIC	Mallow's Cp	CV PRESS
With central obesity and waist%	54	0.4467	0.3181	-91.75	-138.40	8.15	3.48
Without central obesity and waist%	75	0.2075	0.1740	-137.22	-211.35	0.41	4.36

Table 36. Model selection criteria without central obesity and waist% for non-missing dataset for serum uric acid model

	N	R^2	Adjusted R^2	AIC	BIC	Mallow's Cp	CV PRESS
Without central obesity and waist% for non-missing dataset	54	0.3447	0.2765	-92.62	-144.10	2.37	3.45

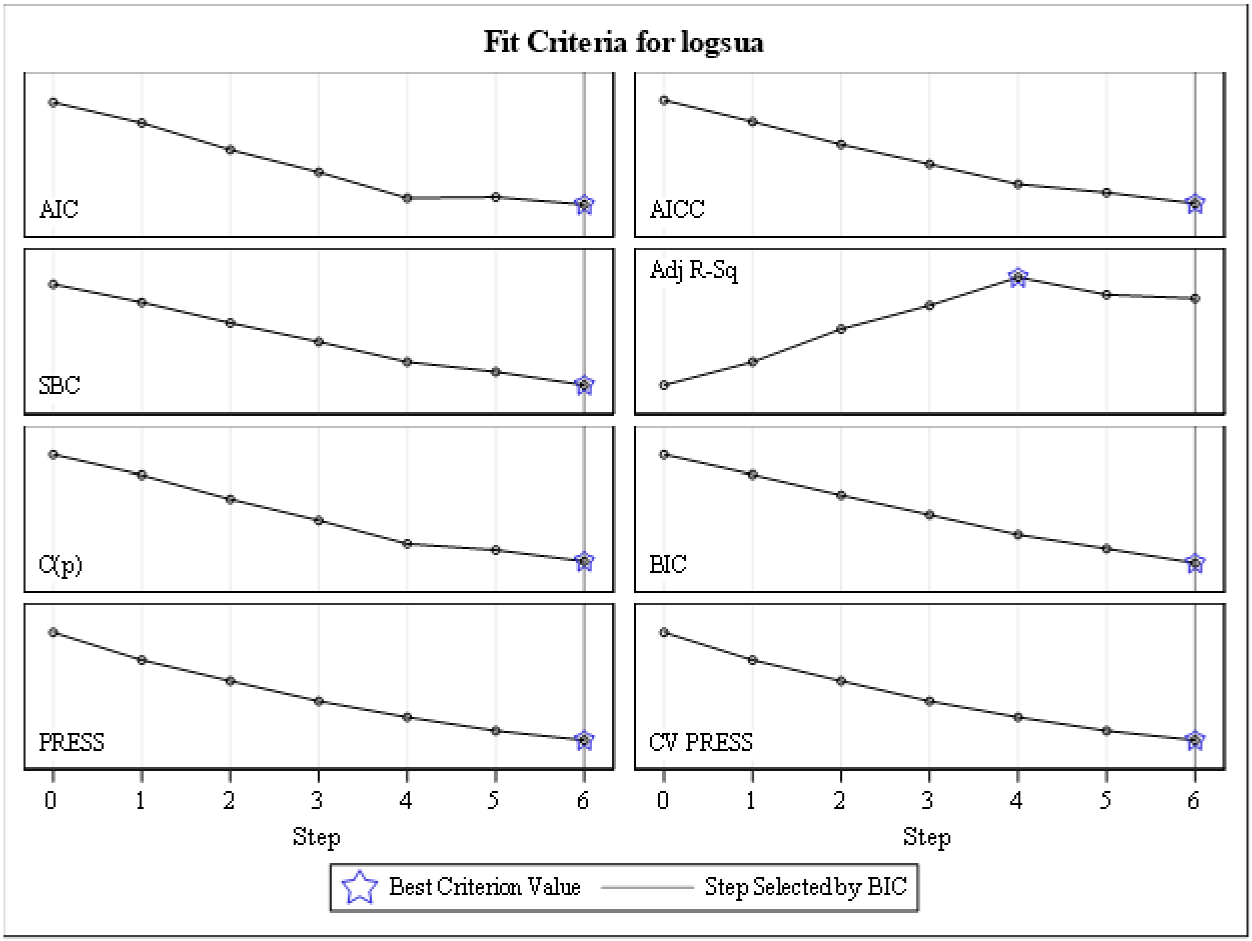
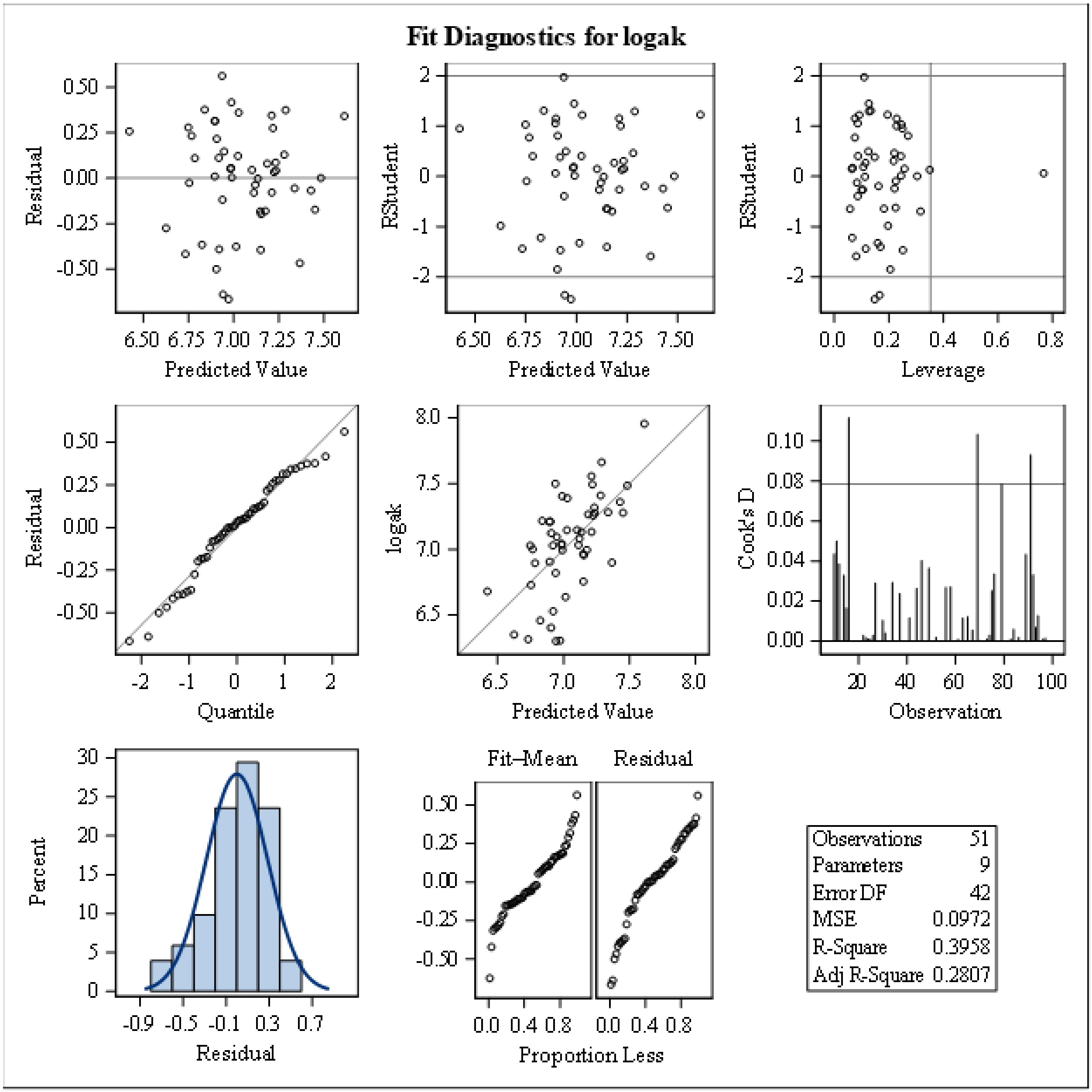


Figure 14. Fit criteria for the best selected log (serum uric acid) model



Observations	51
Parameters	9
Error DF	42
MSE	0.0972
R-Square	0.3958
Adj R-Square	0.2807

Figure 15. Regression diagnostics for log (α-klotho) model

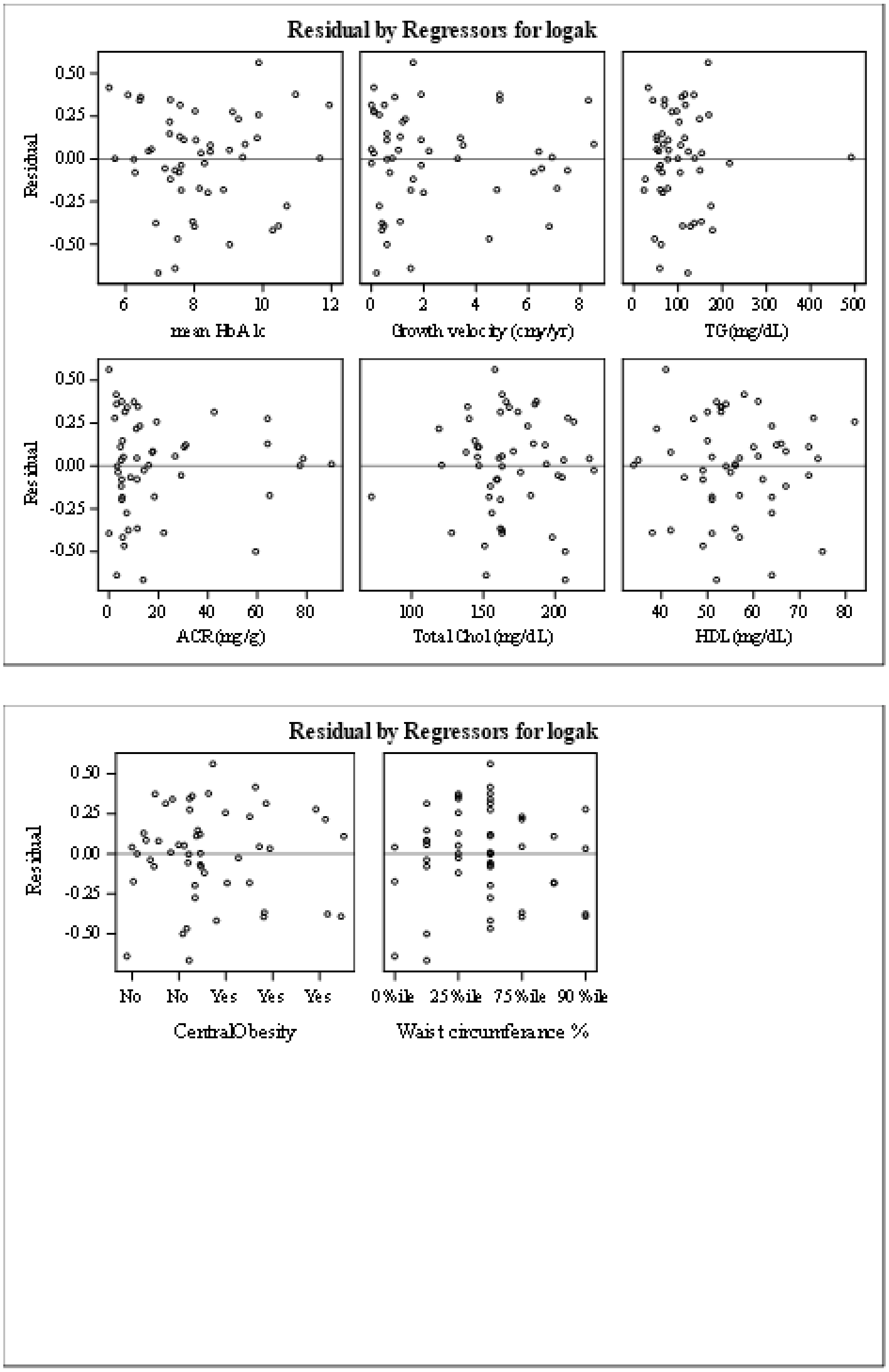
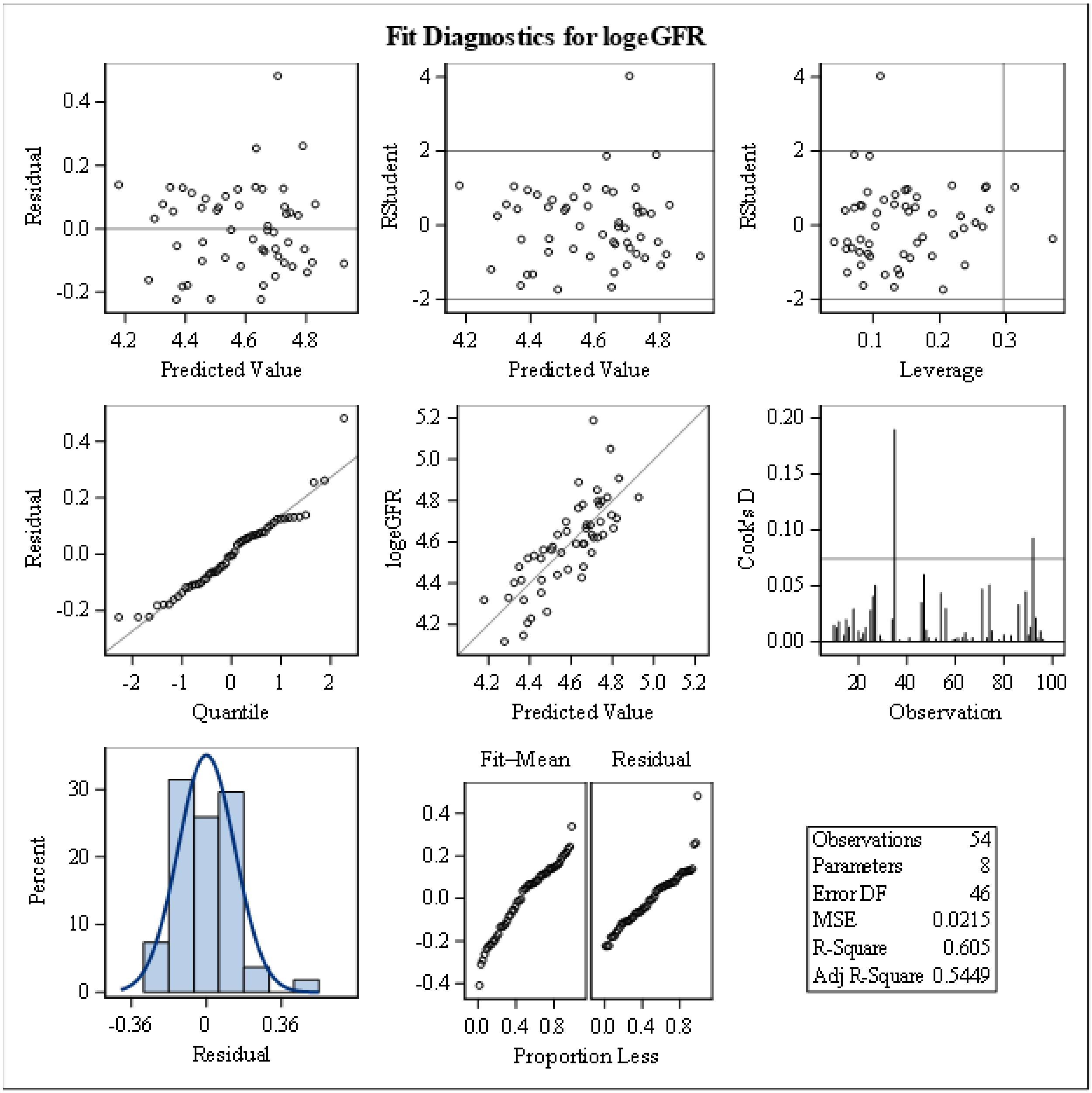


Figure 16. Residual analysis for log (α-klotho) model



Observations	54
Parameters	8
Error DF	46
MSE	0.0215
R-Square	0.605
Adj R-Square	0.5449

Figure 17. Regression diagnostics for log (eGFR) model

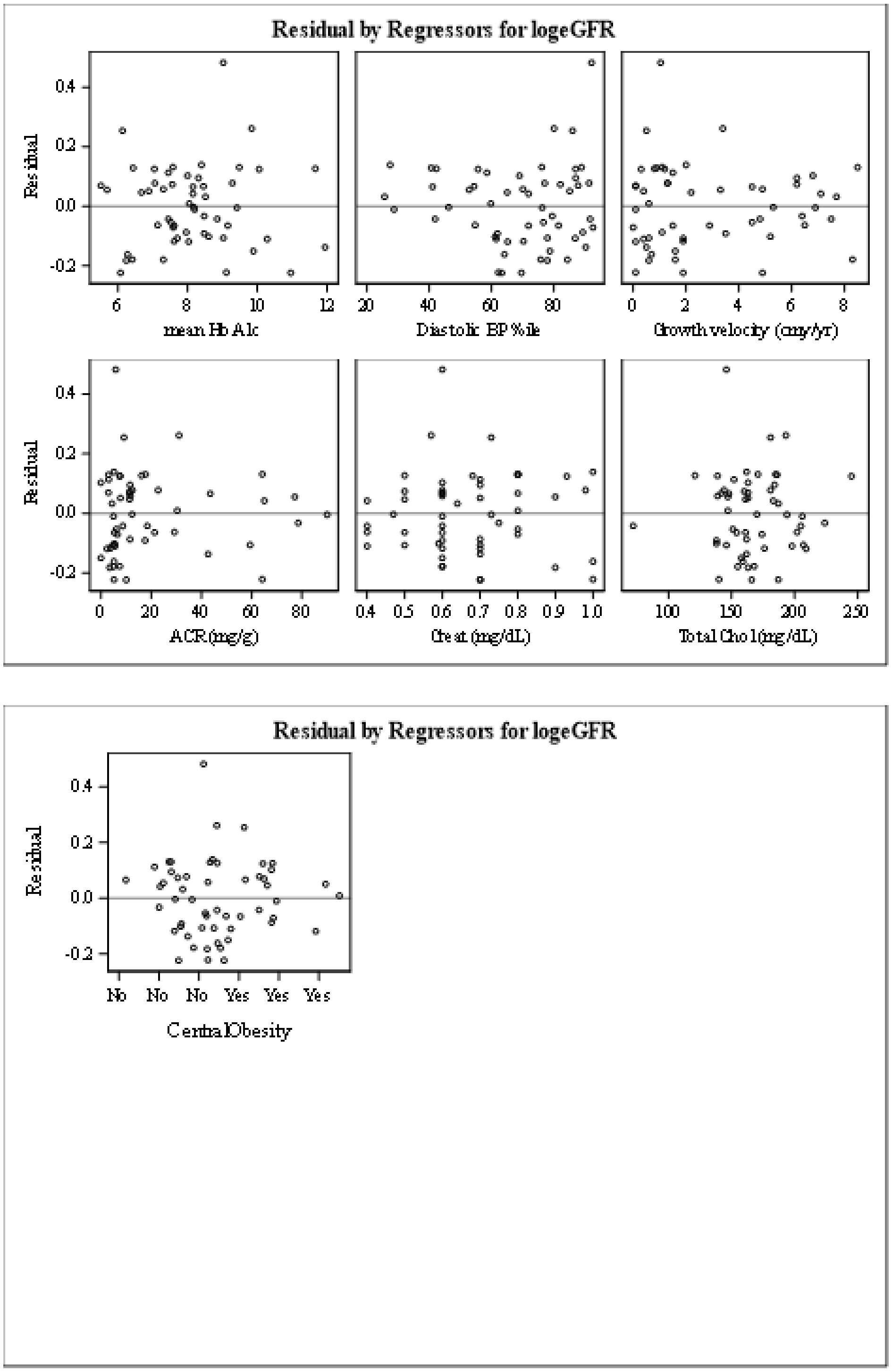
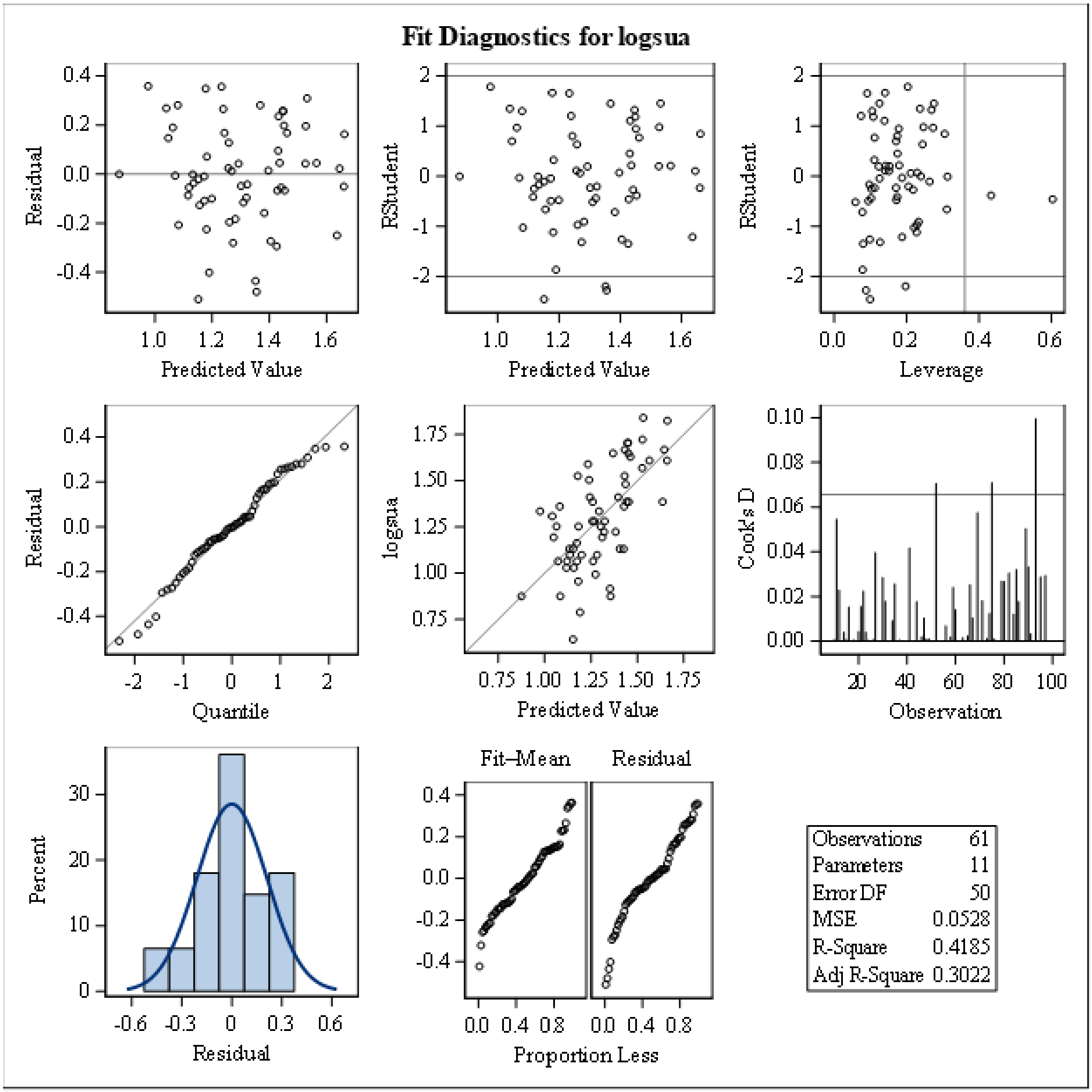


Figure 18. Residual analysis for log (eGFR) model



Observations	61
Parameters	11
Error DF	50
MSE	0.0528
R-Square	0.4185
Adj R-Square	0.3022

Figure 19. Regression diagnostics for log (serum uric acid) model

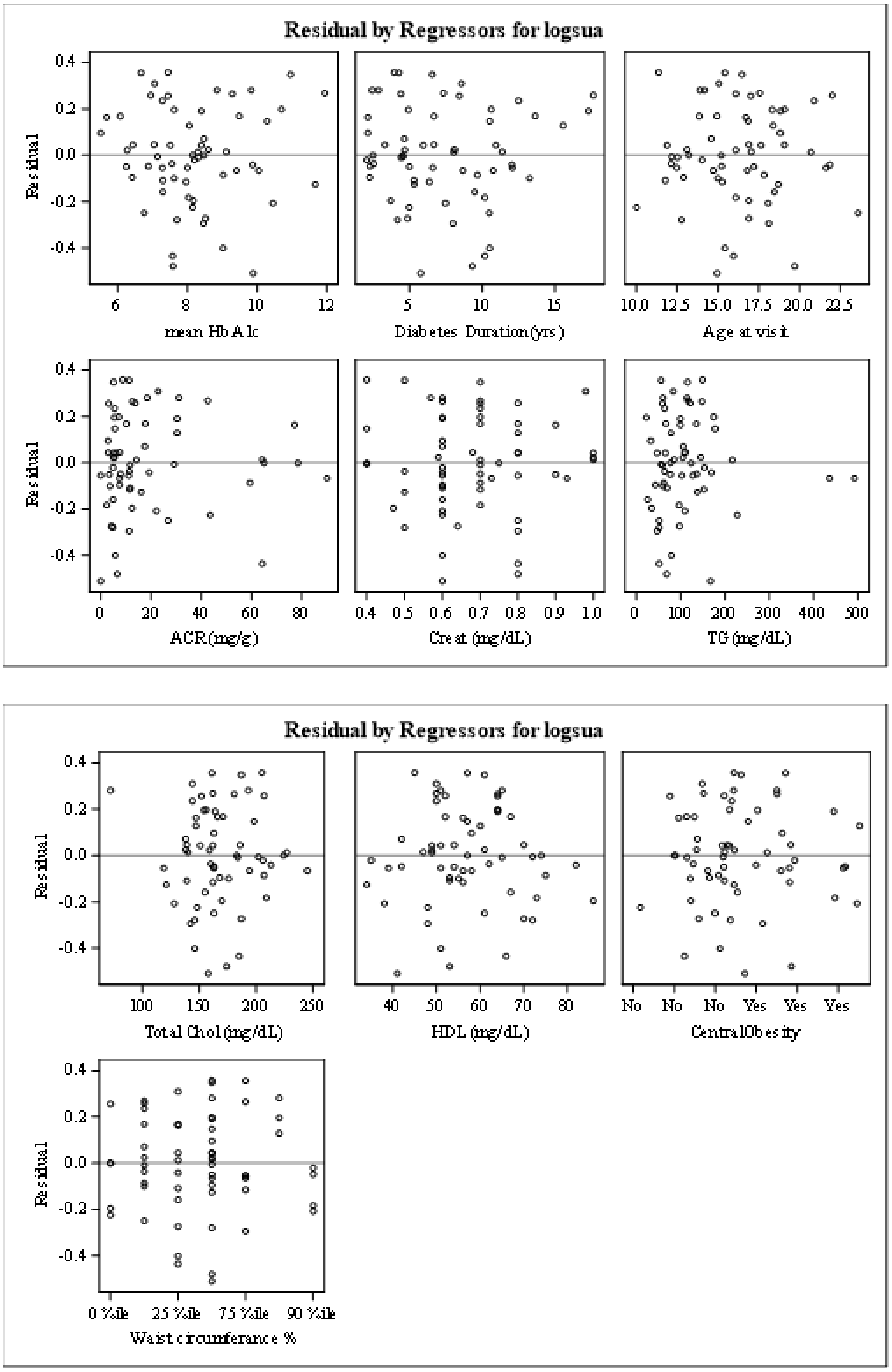


Figure 20. Residual analysis for log (serum uric acid) model

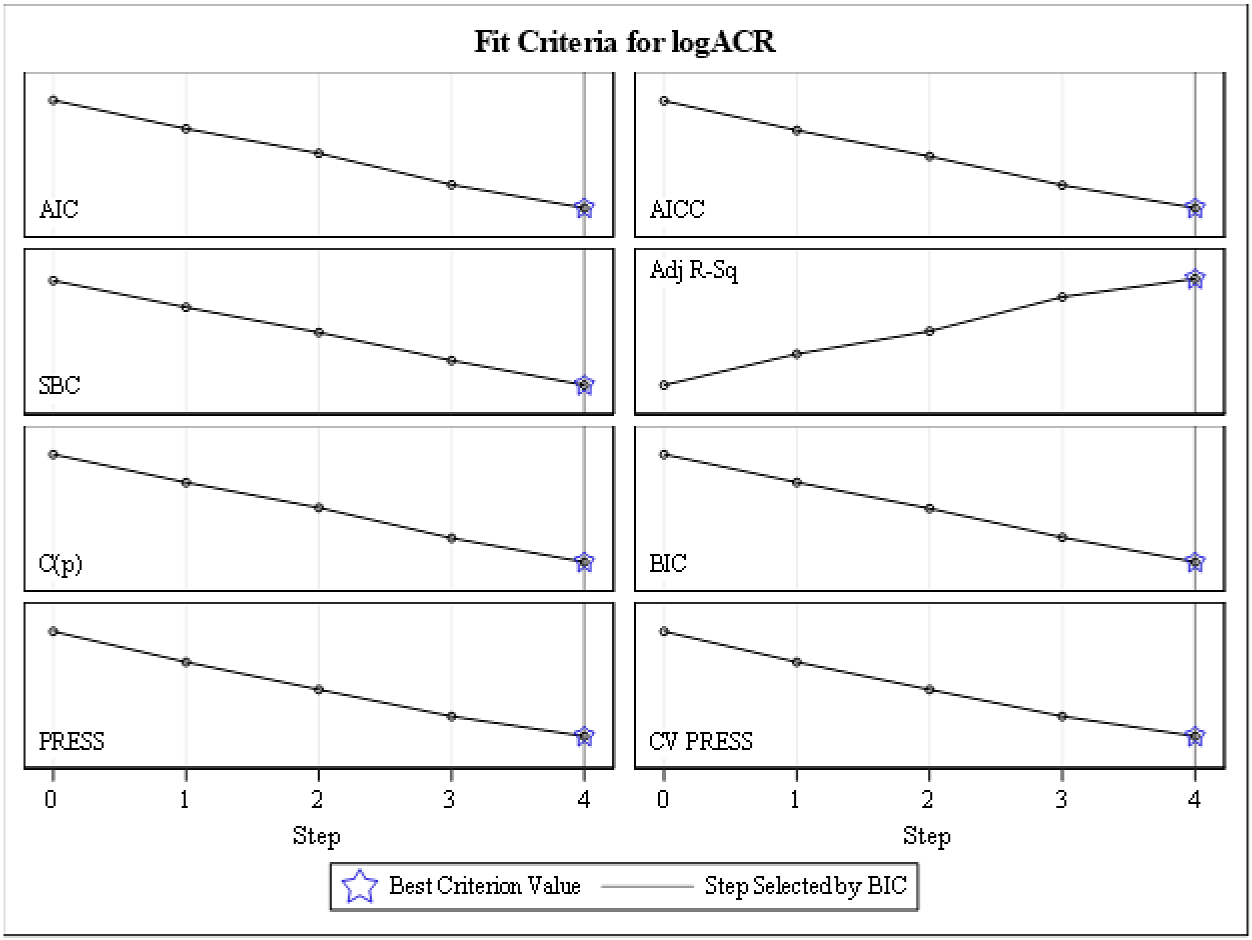


Figure 21. Fit criteria for the log (ACR) model with biomarkers and other variables

Table 37. VIF of log (ACR) model with biomarkers and other variables

	HbA1c	Diabetic duration	Age	SBP%	DBP%	TG	HDL
VIF	1.20	1.24	1.13	1.84	1.88	1.16	1.07

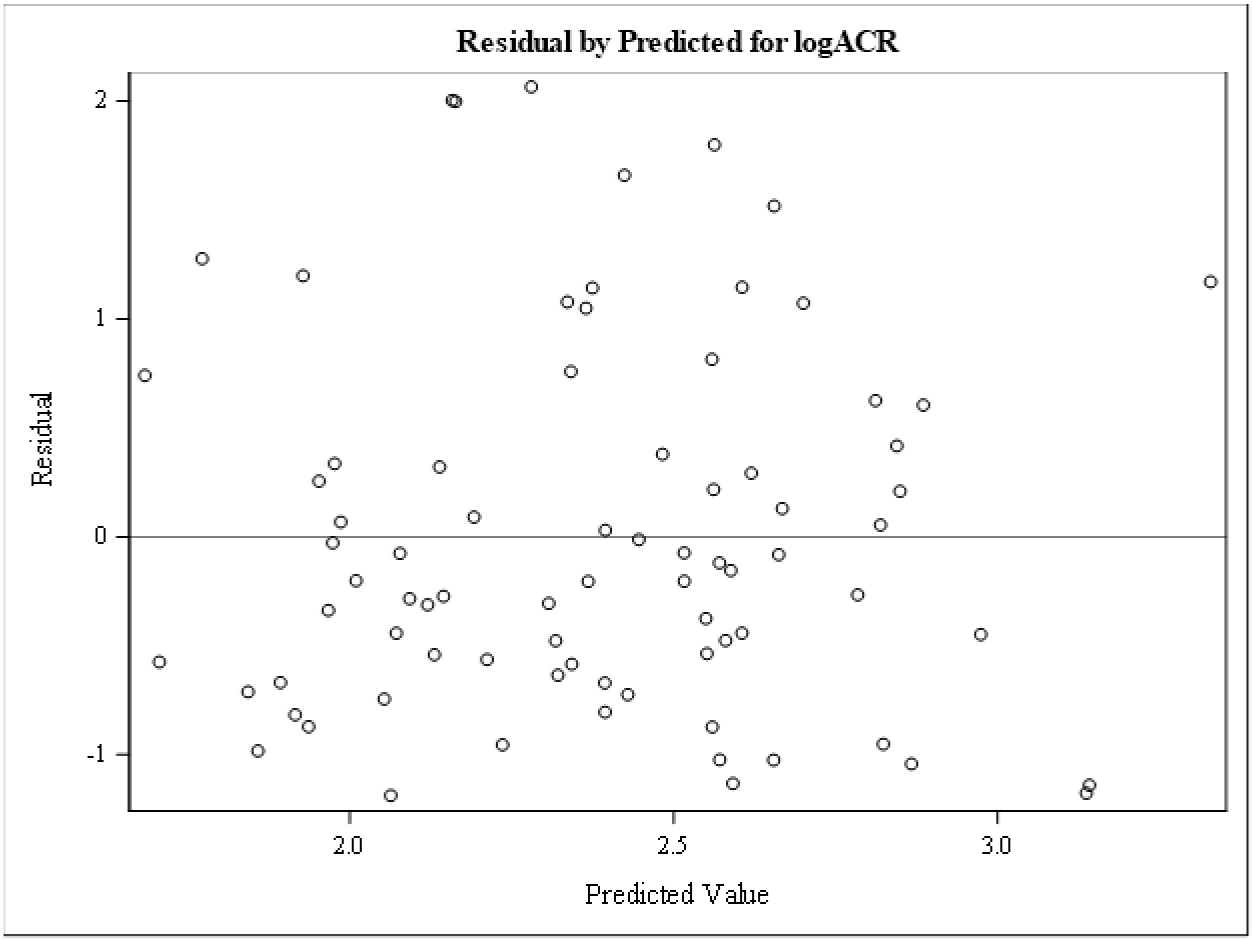


Figure 22. Scatter plot of residuals by predicted values for log (ACR) model

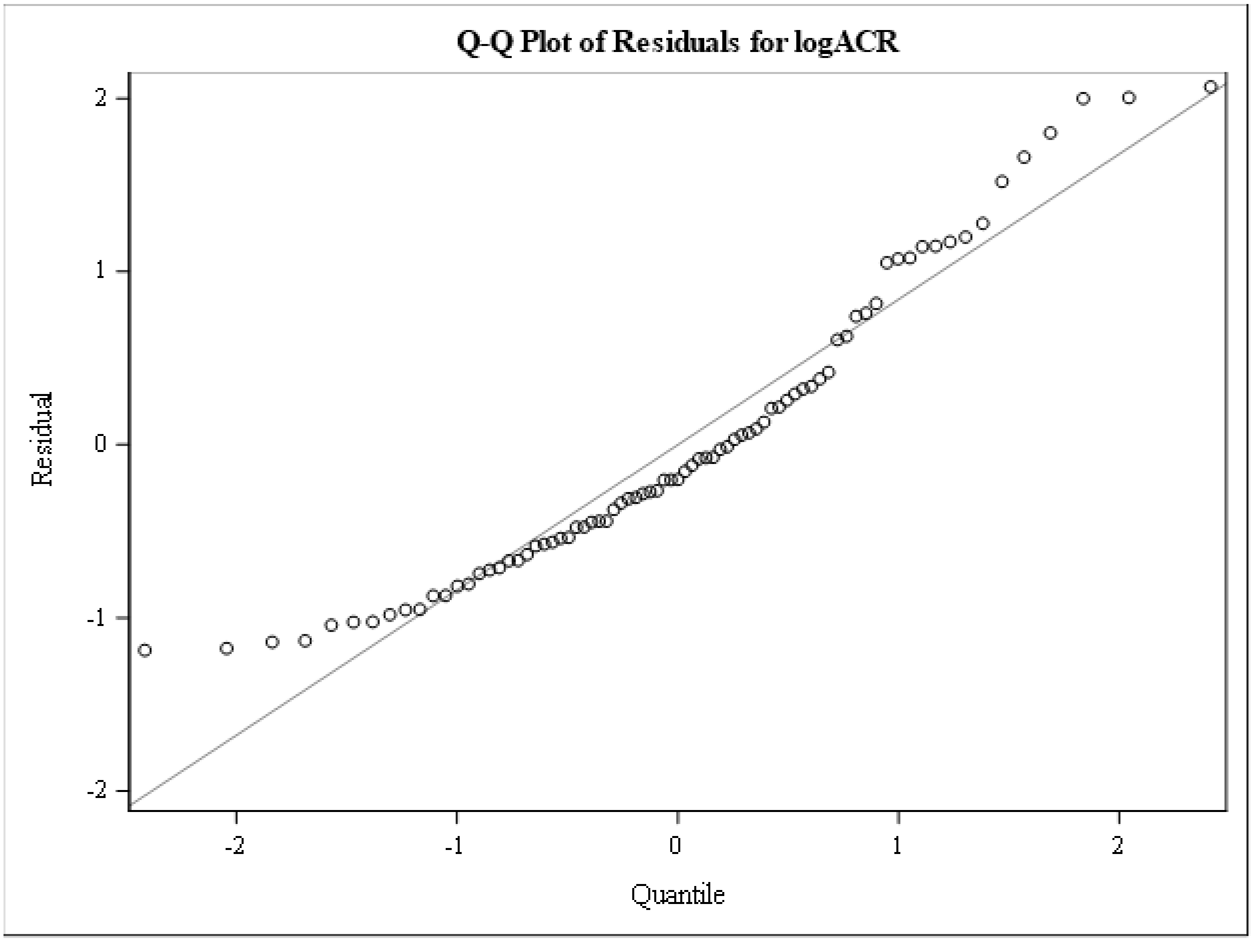


Figure 23. QQ plot of residuals for log (ACR) model

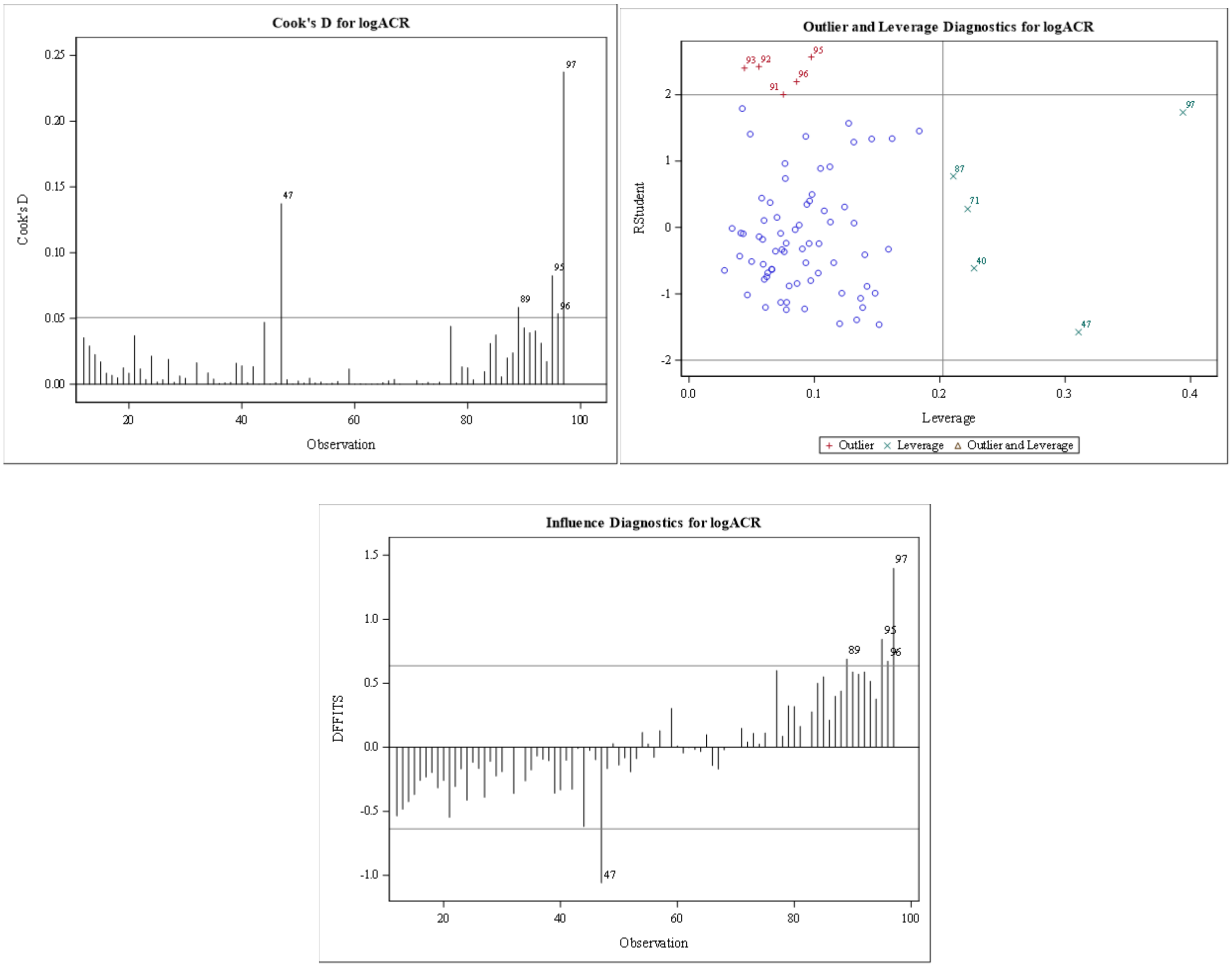
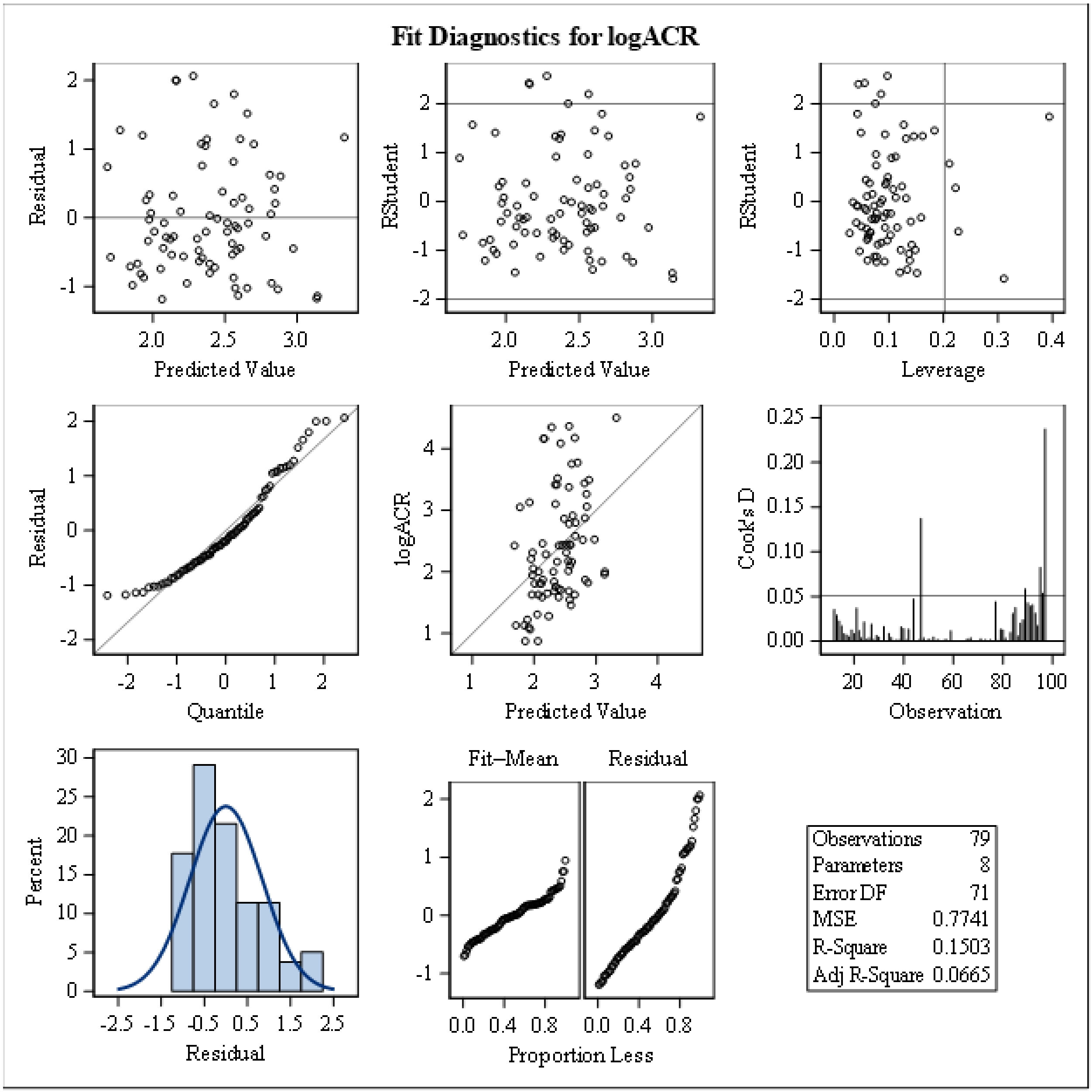


Figure 24. Plots of problematic points for log (ACR) model



Observations	79
Parameters	8
Error DF	71
MSE	0.7741
R-Square	0.1503
Adj R-Square	0.0665

Figure 25. Regression diagnostics for log (ACR) model

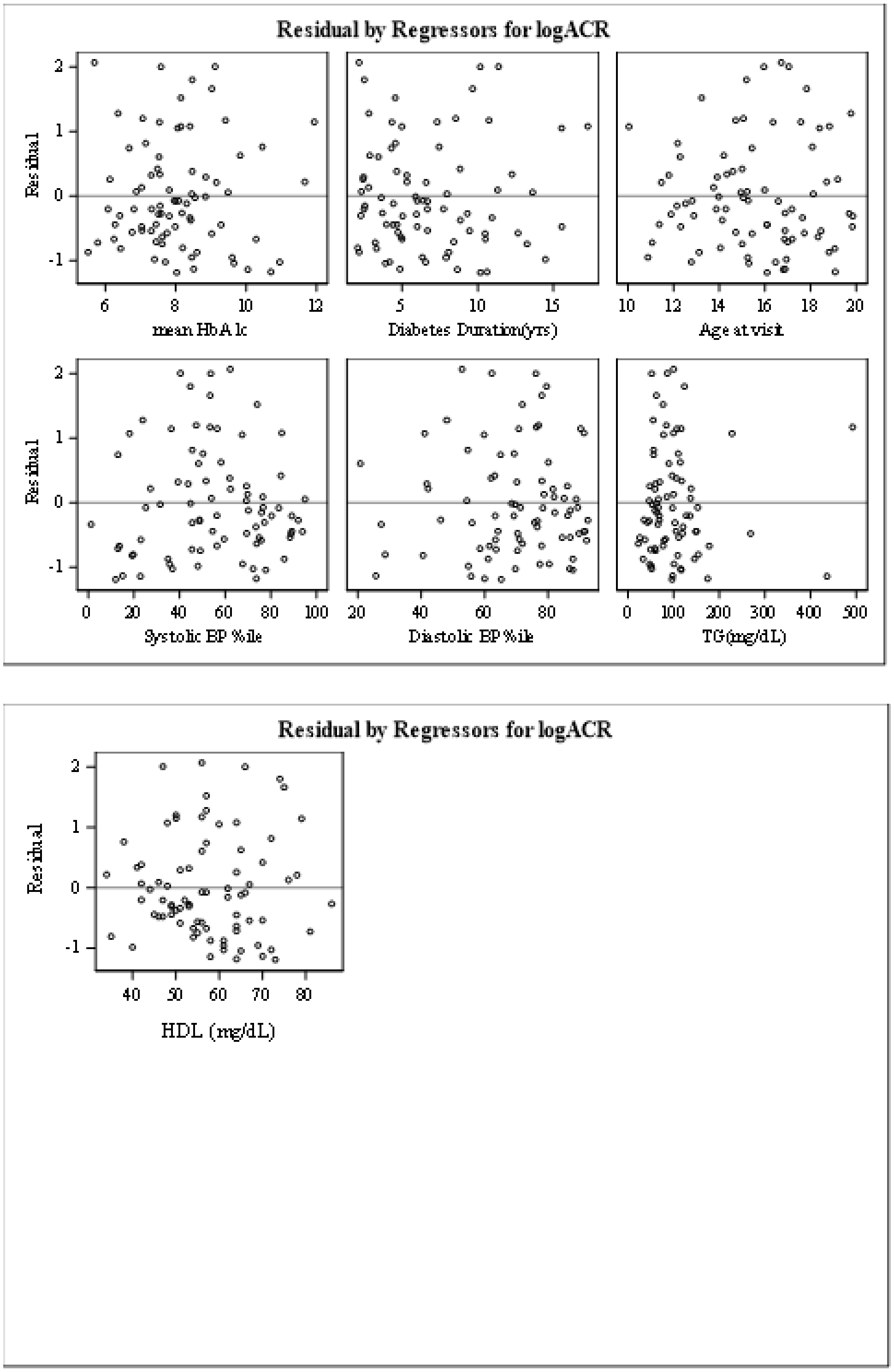


Figure 26. Residual analysis for log (ACR) model

APPENDIX B: RELEVANT SAS CODES

```
LIBNAME results 'C:\Users\fuhaoyi\Desktop\study\thesis\thesis writing';
LIBNAME DN 'C:\Users\fuhaoyi\Desktop\study\thesis\2017 Spring';

PROC FORMAT;
  VALUE include 1='Yes' 0='No';
  VALUE gender 1='Male' 2='Female';
  VALUE racegroup 1='White' 2='AA' 3='Asian' 4='Hispanic' 5='Other';
  VALUE insulin 1='MDI' 2='CSII' 3='Pre-mixed';
  VALUE rscreen 1='CHP' 2='OSH';
  VALUE rstatus 0='Neg' 1='NP' 2='P';
  VALUE medi 0='None' 1='ACE' 2='Other Antihypertensive' 3='Metformin'
4='OCP' 5='Statin';
  VALUE diabN 1='NA' 2='IMA' 3='PMA' 4='MA';
  VALUE diabyrs 0-<5 = '<5' 5-high = '>=5';
  VALUE tanner 1='pre-pubertal' 2-3 = 'early-pubertal' 4-5 = 'late-pubertal';
  VALUE bmi 0-<85 = 'normal' 85-<95 = 'overweight' 95-high = 'obese';
  VALUE bmi2gp 0-<85 = 'normal' 85-high = 'abnormal';
  VALUE ratio 0-<0.5 = 'No' 0.5-high = 'Yes';
  VALUE growth 0-<5 = 'slow' 5-high = 'fast';
  VALUE A1cPOC 0-<7.5 = '<7.5' 7.5-high = '>=7.5';
  VALUE ACR1gp 0-<30 = 'Normal' 30-<300 = 'micro albuminuria' 300-high =
'macro albuminuria';
  VALUE ACR2gp 0-<30 = 'Normal' 30-high = 'Abnormal';
  VALUE AER 0-<30 = 'Normal' 30-<300 = 'micro albuminuria' 300-high = 'macro
albuminuria';
  VALUE quartiles 0 = '1' 1 = '2' 2 = '3' 3 = '4';
  VALUE race2gp 1 = 'White' 2 = 'AA';
  VALUE lwc_perc 0='0 %ile' 1='10 %ile' 2='25 %ile' 3='50 %ile' 4='75 %ile'
5='85 %ile' 6='90 %ile' ;
  VALUE LB_SEX 1=Female 2=Male;

RUN;

data thesis_results;
  set DN.pedrodata_haoyi_2017_03_11;
run;

DATA final_analysis;
  SET thesis_results(keep= mrn AgeCollection DiabDur p_sbp_score p_dbp_score
new_BMIPCT whratio GrowthVel meanA1c ACR
      aklothoSol eGFR SerumUricAcid Creat TG TotalChol LDL HDL sex race
CentralObesity wc_perc A1cPOCgp ACR2gpd);
RUN;

DATA final_analysis;
  SET final_analysis;
    logeGFR=log(eGFR);
      logak=log(aklothoSol);
      logsua=log(SerumUricAcid);
RUN;

DATA final_analysis;
  SET final_analysis;
  IF ACR ne 536.2;
RUN;

data final_analysis;
  set final_analysis;
    if A1cPOCgp = <7.5 then A1cPOC2gp=0;
    if A1cPOCgp > 7.5 then A1cPOC2gp=1;
      if -888<ACR<30 then ACR2gpd2=0;
      if ACR>=30 then ACR2gpd2=1;
      if MISSING(ACR) then ACR2gpd2=.;
      sex2 = sex -1;
      if race =1 then race2=1;
      else if race =2 then race2=0;
      else if race=3 or race =4 or race=5 then race2=.;
      if .<whratio<0.5 then co=0;
      if whratio>=0.5 then co=1;
      if missing(whratio) then co=.;
run;

/* table 1 */
ODS RTF FILE='C:\Users\fuhaoyi\Desktop\study\thesis\thesis
writing\descriptive.rtf'STYLE=journal;
PROC MEANS DATA=final_analysis MAXDEC=2 N MEAN STDDEV P25 MEDIAN P75 MIN MAX;
  VAR AgeCollection DiabDur p_sbp_score p_dbp_score new_BMIPCT whratio
GrowthVel meanA1c ACR
      aklothoSol eGFR SerumUricAcid Creat TG TotalChol LDL HDL;
RUN;

/* table 2 */
PROC FREQ DATA=final_analysis;
  TABLES sex race CentralObesity wc_perc A1cPOCgp ACR2gpd;
RUN;

ODS RTF CLOSE;

%macro table3(varname1, varname2, name);
   ods output spearmancorr=&name;
   proc corr data=final_analysis spearman;
     with &varname1;
       var &varname2;
   run;

   data &name;
     set &name;
     format _numeric_ 5.2;
   run;

   ods rtf file= 'C:\Users\fuhaoyi\Desktop\study\thesis\thesis
writing\correlation.rtf';
   proc print data=&name;
   run;
   ods rtf close;
```

```
%mend;

/* Table 3 Correlations between alpha-klotho and other independent variables
*/
%table3 (AgeCollection DiabDur p_sbp_score p_dbp_score new_BMIPCT whratio
GrowthVel meanA1c ACR
          Creat TG TotalChol LDL HDL,aklothoSol, correlation_ak);

/* Table 4 Correlations between eGFR and other independent variables */
%table3 (AgeCollection DiabDur p_sbp_score p_dbp_score new_BMIPCT whratio
GrowthVel meanA1c ACR
          Creat TG TotalChol LDL HDL, eGFR, correlation_eGFR);

/* Table 5 Correlations between serum uric acid and other independent
variables */
%table3 (AgeCollection DiabDur p_sbp_score p_dbp_score new_BMIPCT whratio
GrowthVel meanA1c ACR
          Creat TG TotalChol LDL HDL,SerumUricAcid, correlation_sua);

PROC CORR DATA=final_analysis;
  VAR AgeCollection DiabDur p_sbp_score p_dbp_score new_BMIPCT whratio
GrowthVel meanA1c ACR
          Creat TG TotalChol LDL HDL;
RUN;

%macro uniak(varname1);
   proc reg data=final_analysis plots=none;
     model logak=&varname1;
   run;
%mend;

/*Table 6. Univariable regression models for log a-klotho */
ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis writing\uniak.rtf';
%uniak (AgeCollection); %uniak (DiabDur);  %uniak (p_sbp_score);  %uniak
(p_dbp_score); %uniak (new_BMIPCT);
%uniak (whratio); %uniak (GrowthVel);  %uniak (meanA1c); %uniak (ACR); %uniak
(Creat);
%uniak (TG); %uniak (TotalChol); %uniak (LDL); %unieGFR (HDL); %uniak (sex2);
%uniak (race2); %uniak (co); %uniak (wc_perc); %uniak (A1cPOC2gp); %uniak
(ACR2gpd2);
ods rtf close;

%macro unieGFR(varname1);
   proc reg data=final_analysis plots=none;
     model logeGFR=&varname1;
   run;
%mend;

/* Table 7. Univariable regression models for logeGFR */
ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis
writing\unieGFR.rtf';
%unieGFR (AgeCollection); %unieGFR (DiabDur);  %unieGFR
(p_sbp_score);  %unieGFR (p_dbp_score); %unieGFR (new_BMIPCT);
%unieGFR (whratio); %unieGFR (GrowthVel);  %unieGFR (meanA1c); %unieGFR
(ACR); %unieGFR (Creat);
%unieGFR (TG); %unieGFR (TotalChol); %unieGFR (LDL); %unieGFR (HDL); %unieGFR
(sex2);
```

```
%unieGFR (race2); %unieGFR (co); %unieGFR (wc_perc); %unieGFR
(A1cPOC2gp); %unieGFR (ACR2gpd2);
ods rtf close;

%macro unisua(varname1);
   proc reg data=final_analysis plots=none;
     model logsua=&varname1;
   run;
%mend;

/* Table 8. Univariable regression models for log serum uric acid */
ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis
writing\unisua.rtf';
%unisua (AgeCollection); %unisua (DiabDur);  %unisua (p_sbp_score);  %unisua
(p_dbp_score); %unisua (new_BMIPCT);
%unisua (whratio); %unisua (GrowthVel);  %unisua (meanA1c); %unisua
(ACR); %unisua (Creat);
%unisua (TG); %unisua (TotalChol); %unisua (LDL); %unisua (HDL); %unisua
(sex2);
%unisua (race2); %unisua (co); %unisua (wc_perc); %unisua
(A1cPOC2gp); %unisua (ACR2gpd2);
ods rtf close;

PROC CORR DATA=final_analysis;
  VAR AgeCollection DiabDur p_sbp_score p_dbp_score new_BMIPCT whratio
GrowthVel meanA1c ACR
          Creat TG TotalChol LDL HDL;
RUN;

/* Table 9. Selected variables for different model selection criteria and k-
fold cross-validation for log(alpha-klotho) */
proc reg data=final_analysis;
model logak=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward
slstay=0.2;
run;

data final_analysis2;
  set final_analysis;
  if centralobesity ne .;
run;

data final_analysis2;
  set final_analysis2;
  if wc_perc ne .;
run;

/* log(alpha-klotho) model with centralobesity and waist% */
/* no cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logak=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=press)
```

```
cvMethod=split(2) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* 2-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logak=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(2) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* 5-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logak=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(5) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* 10-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logak=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(10) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* n-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logak=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(45) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* log(alpha-klotho) model without centralobesity and waist% */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logak=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 /selection=backward (choose=BIC
select=cv)
cvMethod=split(66) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* log(alpha-klotho) model without centralobesity and waist% excluded missing
data */
proc glmselect data=final_analysis2
plots(stepAxis=number)=(criterionPanel ASEPlot);
```

```
model logak=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 /selection=backward (choose=BIC
select=cv)
cvMethod=split(45) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* n-fold cross-validation */
ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis writing\ak fit
stat.rtf';
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logak=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(45) stats=all cvdetails=cvpress SHOWPVALS;
run;
ods rtf close;


/* log(eGFR) model with central obesity and waist percentile*/
/* no cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logeGFR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score
new_BMIPCT GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=press)
cvMethod=split(2) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* 2-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logeGFR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score
new_BMIPCT GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(2) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* 5-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logeGFR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score
new_BMIPCT GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(5) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* 10-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logeGFR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score
new_BMIPCT GrowthVel ACR
```

```
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(10) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* n-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logeGFR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score
new_BMIPCT GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(52) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* log(eGFR) model without centralobesity and waist% */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logeGFR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score
new_BMIPCT GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2/selection=backward (choose=BIC
select=cv)
cvMethod=split(70) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* log(eGFR) model without centralobesity and waist% excluded missing data */
proc glmselect data=final_analysis2
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logeGFR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score
new_BMIPCT GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2/selection=backward (choose=BIC
select=cv)
cvMethod=split(52) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* n-fold cross-validation */
ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis writing\eGFR fit
stat.rtf';
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logeGFR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score
new_BMIPCT GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(52) stats=all cvdetails=cvpress SHOWPVALS;
run;
ods rtf close;

/* log(serum uric acid) model with central obesity and waist percentile*/
/* no cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logsua=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity wc_perc
/selection=backward (choose=BIC select=press)
cvMethod=split(2) stats=all cvdetails=cvpress SHOWPVALS details=all;
```

```
run;

/* 2-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logsua=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(2) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* 5-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logsua=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(5) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* 10-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logsua=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(10) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* n-fold cross-validation */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logsua=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(54) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* log(serum uric acid) model without centralobesity and waist% */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logsua=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 /selection=backward (choose=BIC
select=cv)
cvMethod=split(75) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* log(serum uric acid) model without centralobesity and waist% excluded
missing data */
proc glmselect data=final_analysis2
plots(stepAxis=number)=(criterionPanel ASEPlot);
```

```
model logsua=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 /selection=backward (choose=BIC
select=cv)
cvMethod=split(54) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* n-fold cross-validation */
ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis writing\sua fit
stat.rtf';
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logsua=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score new_BMIPCT
GrowthVel ACR
          Creat TG TotalChol HDL sex2 race2 centralobesity
wc_perc/selection=backward (choose=BIC select=cv)
cvMethod=split(54) stats=all cvdetails=cvpress SHOWPVALS;
run;
ods rtf close;

ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis writing\vif.rtf';
/* vif checking for logak */
proc reg data=final_analysis;
  model logak=meanA1c growthVel TG ACR TotalChol HDL Centralobesity wc_perc /
vif;
run;

/* vif checking for logeGFR */
proc reg data=final_analysis;
  model logeGFR=meanA1c p_dbp_score growthVel ACR Creat TotalChol
Centralobesity / vif;
run;

/* vif checking for logsua */
proc reg data=final_analysis;
  model logsua=meanA1c DiabDur AgeCollection ACR Creat TG TotalChol HDL
Centralobesity wc_perc/ vif;
run;
ods rtf close;

ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis
writing\assumption.rtf';
/* linearity checking for logak */
proc reg data=final_analysis
  plots=(RESIDUALBYPREDICTED);
  model logak=meanA1c growthVel TG ACR TotalChol HDL Centralobesity wc_perc;
run;

/* linearity checking for logeGFR */
proc reg data=final_analysis
  plots=(RESIDUALBYPREDICTED);
  model logeGFR=meanA1c p_dbp_score growthVel ACR Creat TotalChol
Centralobesity;
run;

/* linearity checking for logsua */
proc reg data=final_analysis
```

```
  plots=(RESIDUALBYPREDICTED);
  model logsua=meanA1c DiabDur AgeCollection ACR Creat TG TotalChol HDL
Centralobesity wc_perc;
run;

/* Normality checking for logak */
proc reg data=final_analysis
  plots=(QQPLOT);
  model logak=meanA1c growthVel TG ACR TotalChol HDL Centralobesity wc_perc;
run;

/* Normality checking for logeGFR */
proc reg data=final_analysis
  plots=(QQPLOT);
  model logeGFR=meanA1c p_dbp_score growthVel ACR Creat TotalChol
Centralobesity;
run;

/* Normality checking for logsua */
proc reg data=final_analysis
  plots=(QQPLOT);
  model logsua=meanA1c DiabDur AgeCollection ACR Creat TG TotalChol HDL
Centralobesity wc_perc;
run;
ods rtf close;

ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis
writing\problematic points.rtf ';
/* Problematic points for logak */
proc reg data=final_analysis
plots=(RStudentByLeverage(label) CooksD(label) DFFITS(label));
model logak=meanA1c growthVel TG ACR TotalChol HDL Centralobesity wc_perc;
run;
ods rtf close;

/* Problematic points for logeGFR */
proc reg data=final_analysis
plots=(RStudentByLeverage(label) CooksD(label) DFFITS(label));
model logeGFR=meanA1c p_dbp_score growthVel ACR Creat TotalChol
Centralobesity;
run;

/* Problematic points for logsua */
proc reg data=final_analysis
plots=(RStudentByLeverage(label) CooksD(label) DFFITS(label));
model logsua=meanA1c DiabDur AgeCollection ACR Creat TG TotalChol HDL
Centralobesity wc_perc;
run;
ods rtf close;

ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis writing\ACR.rtf';
/* ACR normality check */
proc univariate data=final_analysis;
  var ACR;
  histogram ACR / normal;
run;
```

```
/* logACR */
data final_analysis;
  set final_analysis;
  logACR=log(ACR);
run;

/* log(ACR) normality check */
proc univariate data=final_analysis;
  var logACR;
  histogram logACR / normal;
run;

/* logACR model with three biomarkers */
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logACR=meanA1c DiabDur AgeCollection aklothoSol eGFR SerumUricAcid
/selection=backward (choose=BIC select=cv include=3)
cvMethod=split(66) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* logACR model with three biomarkers and other variables*/
proc glmselect data=final_analysis
plots(stepAxis=number)=(criterionPanel ASEPlot);
model logACR=meanA1c DiabDur AgeCollection aklothoSol eGFR SerumUricAcid
             p_sbp_score p_dbp_score TG HDL LDL/selection=backward
(choose=BIC select=cv include=3)
cvMethod=split(61) stats=all cvdetails=cvpress SHOWPVALS details=all;
run;

/* vif checking for logACR */
proc reg data=final_analysis;
  model logACR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score TG HDL /
vif;
run;

/* linearity checking for logACR */
proc reg data=final_analysis
  plots=(RESIDUALBYPREDICTED);
  model logACR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score TG HDL;
run;

/* Normality checking for logACR */
proc reg data=final_analysis
  plots=(QQPLOT);
  model logACR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score TG HDL;
run;
ods rtf file='C:\Users\fuhaoyi\Desktop\study\thesis\thesis writing\ACR.rtf';
/* Problematic points for logACR */
proc reg data=final_analysis
plots=(RStudentByLeverage(label) CooksD(label) DFFITS(label));
model logACR=meanA1c DiabDur AgeCollection p_sbp_score p_dbp_score TG HDL;
run;
ods rtf close;
```

BIBLIOGRAPHY

[1]. Zysberg L, Yoseph TB, et al., *Emotional intelligence and glycemic management among type I diabetes patients.* Journal of Health Psychology, 2017.22(2): 158-163.

[2]. American Diabetes Association (2014) *Infographic: A Snapshot of Diabetes in America*. Available at: http://www.diabetes.org/diabetes-basics/statistics/cdc-infographic.html

[3]. World Health Organization (WHO) (2017) *Global report on diabetes*. Available at: http://who.int/diabetes/global-report/en/

[4]. American Diabetes Association, *Diagnosis and classification of diabetes mellitus*. Diabetes Care, 2009.32(Suppl 1): S62-67.

[5]. The DIAMOND Project Group, *Incidence and trend of childhood type 1 diabetes worldwide 1990-1999*. Diabet Med, 2006.23:857-866.

[6]. Gan MJ, O'Neill AA, Haller MJ, *Type 1 Diabetes: Current Concepts in Epidemiology, Pathophysiology, Clinical Care, and Research.* Current Problems in Pediatric and Adolescent Health Care, 2012.42(10):269-291.

[7]. American Diabetes Association, *Diagnosis and classification of diabetes mellitus*. Diabetes Care, 2009.32(Suppl 1): S62-67.

[8]. The DIAMOND Project Group, *Incidence and trend of childhood type 1 diabetes worldwide 1990-1999*. Diabet Med, 2006.23:857-866.

[9]. Center for Disease Control (CDC) (2014) *National Diabetes Statistics Report, 2014*. Available at: https://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf

[10]. Mogensen CE, et al., *Prevention of diabetic renal disease with special reference to microalbuminuria*. Lancet, 1995.346:1080-1084.

[11]. Schultz CJ, et al., *Microalbuminuria prevalence varies with age, sex, and puberty in children with type 1 diabetes followed from diagnosis in a longitudinal study*. Oxford Regional Prospective Study Group, Diabetes Care, 1999.22:495-502.

[12]. Groop PH, Thomas MC, Moran JL, et al., *The presence and severity of chronic kidney disease predicts all-cause mortality in type 1 diabetes*. *Diabetes*, 2009.58:1651-1658.

[13]. Donaghue KC, Wadwa RP, Dimeglio LA, et al., *Microvascular and macrovascular complications in children and adolescents*. Pediatric Diabetes, 2014.15(Suppl. 20):257-269.

[14]. Lee SY, Choi ME. *Urinary biomarkers for early diabetic nephropathy: beyond albuminuria*. Pediatr Nephrol, 2015.30(7):1063-75.

[15]. Zachwieja J, Soltysiak J, Fichna P, Lipkowska K, Stankiewicz W, Skowronska B, Kroll P, Lewandowska-Stachowiak M, *Normal-range albuminuria does not exclude nephropathy in diabetic children*. Pediatr Nephrol, 2010. 25(8):1445-51.

[16]. Asai O, Nakatani K, Tanaka T, Sakan H, Imura A, Yoshimoto S, Samejima K, Yamaguchi Y, Matsui M, Akai Y, Konishi N, Iwano M, Nabeshima Y, Saito Y, *Decreased renal α-Klotho expression in early diabetic nephropathy in humans and mice and its possible role in urinary calcium excretion*. Kidney Int, 2012. 81(6):539-47.

[17]. Lee EY, Kim SS, Lee JS, Kim IJ, Song SH, Cha SK, Park KS, Kang JS, Chung CH, *Soluble α-klotho as a novel biomarker in the early stage of nephropathy in patients with type 2 diabetes*. PLoS One, 2014. 9(8):875.

[18]. Rotondi S, Pasquali M, Tartaglione L, et al., *Soluble alpha -Klotho Serum Levels in Chronic Kidney Disease*. Int J Endocrinol, 2011. 6:1599-1608.

[19]. Maltese G, Fountoulakis N, Siow RC, Gnudi L, Karalliedde J, *Perturbations of the anti-ageing hormone Klotho in patients with type 1 diabetes and microalbuminuria*. Diabetologia, 2017.

[20]. Sharma AP, Yasin A, et al., *Diagnostic Accuracy of Cystatin C-Based eGFR Equations at Different GFR Levels in Children*. Clinical Journal of the American Society of Nephrology, 2011.6:1599-1608.

[21]. Mise K, Hoshino J, et al., *Clinical and pathological predictors of estimated GFR decline in patients with type 2 diabetes and overt proteinuric diabetic nephropathy*. Diabetes Metab Res Rev, 2015. 31:572-581.

[22]. Bjornstad P, Cherney D, Maahs DM, *Early diabetic nephropathy in type 1 diabetes: new insights*. Curr Opin Endocrinol Diabetes Obes, 2014. 21(4):279-286.

[23]. Bhole V, Choi JW, Kim SW, et al., *Serum uric acid levels and the risk of type 2 diabetes: A prospective study*. The American Journal of Medicine, 2010.123(10): 957-961.

[24]. Kanbay M, Yilmaz MI, Sonmez A, et al., *Serum uric acid independently predicts cardiovascular events in advanced nephropathy*. American Journal of Nephrology, 2012. 36(4): 324-331.

[25]. Ficociello LH, Rosolowsky ET, et al., *High-normal serum uric acid increases risk of early progressive renal function loss in type 1 diabetes: Results of a 6-year follow-up.* Diabetes Care, 2010.33(6): 1337-1343.

[26]. Bjornstad P, Snell-Bergeon JK, et al., *Serum uric acid and insulin sensitivity in adolescents and adults with and without type 1 diabetes*. Journal of Diabetes and Its Complications, 2014. 28: 298-304.

[27]. Rosner B, Cook N, et al., *Determination of Blood Pressure Percentiles in Normal-Weight Children: Some Methodological Issues*. Am J Epidemiology, 2008. 167(6): 653-666.

[28]. Countinho T, Goel, K, et al., *Central Obesity and Survival in Subjects with Coronary Artery Disease*. Journal of the American College of Cardiology, 2011. 57(19):1877-1886.

[29]. Mushtaq MU, Gull S, et al., *Waist circumference, waist-hip ratio and waist-height ratio percentiles and central obesity among Pakistani children aged five to twelve years*. BMC Pediatrics, 2011. 11:105.

[30]. Cukierman-Yaffe T, Gerstein HC, Williamson JD, et al., *Relationship between baseline glycemic control and cognitive function in individuals with type 2 diabetes and other cardiovascular risk factors: the action to control cardiovascular risk in*

diabetes-memory in diabetes (ACCORD-MIND) trial. Diabetes Care, 2009. 32:221–226.

[31]. Methven S, MacGregor MS, et al., *Assessing proteinuria in chronic kidney disease: protein–creatinine ratio versus albumin–creatinine ratio*. Nephrol Dial Transplant, 2010. 25(9): 2991-2996.

[32]. Nelson, D. L. Cox, M. M. (2000), Lehninger, Principles of Biochemistry (3rd ed.). New York: Worth Publishing.

[33]. "Boston scientists say triglycerides play key role in heart health". The Boston Globe. Retrieved 2014-06-18.

[34]. Eckardstein AV, Nofer JR, et al., *High Density Lipoproteins and Arteriosclerosis Role of Cholesterol Efflux and Reverse Cholesterol Transport*. Arterioscler Thromb Vasc Biol, 2001. 21:13-27.

[35]. Ai Masumi, Otokozawa S, et al., *Small dense LDL Cholesterol and coronary heart disease: Results from the Framingham Offspring Study*. Clinical Chemistry, 2010. 56(6):967-976.

[36]. Chatterjee, S., & Hadi, A. S. (2013). *Regression Analysis by Example*. Somerset: Wiley.

[37]. Tarpey Thaddeus, *A Note on the Prediction Sum of Squares Statistic for Restricted Least Squares*. The American Statistician, 2000. 54(2): 116–118.

[38]. Kohavi, Ron (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2 (12): 1137–1143.

[39]. Grubbs, FE., *Procedures for detecting outlying observations in samples*. Technometrics, 2011. 11(1):1-21.