Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Xinhang Wang Date

Machine learning Application in Longitudinal Polycystic Kidney Disease (PKD) Function Prediction

By

Xinhang Wang

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

Xiangqin Cui, PhD
(Thesis Advisor)

_____

Traci Leong, PhD
(Reader)

Machine learning Application in Longitudinal Polycystic Kidney Disease (PKD) Function Prediction

By

Xinhang Wang

B.A., Shanghai University of Finance & Economics, 2018

Thesis Committee Chair: Xiangqin Cui, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2020

Abstract

Machine learning Application in Longitudinal Polycystic Kidney Disease (PKD) Function Prediction

By Xinhang Wang

Background: Autosomal dominant polycystic kidney disease (ADPKD) is one of the most common genetic chronic kidney diseases. The evaluation is based on kidney function represented by estimated glomerular filtration rate (eGFR). The pathological progression of ADPKD is related with genetic factors, demographic and clinical information. Typical pattern of kidney function for ADPKD patients remains in normal range for a long term and followed by a sharp deterioration, making it hard to predict in early stages. The CRISP study monitored the eGFR value and other factors for 242 early stages ADPKD patients longitudinally.

Methods: We evaluated multiple machine learning methods in predicting eGFR values and yearly change of the CRISP cohort. Predictors include variables of demographics, biomarkers, and imaging dataset. Different years of records were used to evaluate the power of historical information. The cohort was divided into subgroups to test the model performance on patients with different kidney function levels. Several expensive predictors were included or excluded in the models, and important predictors were identified in their contribution to the prediction of eGFR and its decline.

Results: The R^2 of machine learning models predicting Year 2 eGFR value were above 0.64 using Year 1 data, while for models predicting eGFR change the R^2 were around 0. When subgrouping patients, the R^2 was largest (0.64) for predicting eGFR value of patients with abnormal kidney function. The R^2 were below 0.47 when predicting Year 6 eGFR value using Year 2 information. In predicting Year 3 eGFR value, adding more years of historical data or health information slightly improved R^2 by 1-3%. Excluding PKD genotype or total kidney volume did not decrease the R^2.

Discussion: Predicting eGFR value using previous year's information is more powerful than prediction eGFR yearly change. The predictive models performed better for patients with abnormal kidney function in subgroup analysis. The prediction power for eGFR values decreased when projecting into the distant future. Including predictors of health information, and previous year eGFR change helped to a small improvement. The expensive predictors of PKD genotype and total kidney volume can be replaced by biomarker variables without affecting the prediction power.

Key Words: ADPKD, Machine Learning, CRISP Cohort

Machine learning Application in Longitudinal Polycystic Kidney Disease (PKD) Function Prediction

By

Xinhang Wang

B.A., Shanghai University of Finance & Economics, 2018

Thesis Committee Chair: Xiangqin Cui, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2020

Introduction

Autosomal dominant polycystic kidney disease is one of the most common genetic renal disorders, primary characteristic is the enlargement of cysts clusters in kidney and irreversible deterioration of renal function (Higashihara, 2012). ADPKD affects over 600,000 people in the United States and 12 million globally (Helal, 2013), with approximately 60% of patients reporting symptoms of acute and chronic pain distributed throughout the body (Bajwa, 2004). The majority of ADPKD patients progress to end-stage renal disease (ESRD), which makes it the fourth leading cause of kidney failure in the United States and worldwide (Collins, 2012).

ADPKD is typically diagnosed by large kidneys with multiple bilateral cysts through imaging techniques, like renal ultrasonography, magnetic resonance imaging and computed tomography (Chebib, 2016). Its progression to ESRD is assessed through changes in serum creatinine levels and the kidney function, which is represented by estimated glomerular filtration rate (Steven, 2006). The cause of ADPKD progression includes genetic and non-genetic factors. It is genetically determined by the mutation of two genes: PKD1, which encodes polycystin 1 (PC-1), and PKD2, which encodes polycystin 2 (PC-2). PKD1 and PKD2 are inherited according to Mendel's law and the mutation of them represents 85% and 15% cases respectively (Moyer, 1994). The genotype of patients can be identified by Sanger sequencing and followed by multiplex-dependent probe amplification (MLPA) with a considerable cost (Eisenberger, 2015). Total kidney volume (TKV) is another important feature for ADPKD progression. Its increasing rate reflects the pathologic processes. A sequential measurement of TKV can serve as a potential indicator of disease progression and treatment efficacy for ADPKD (Grantham, 2016). Some other factors used in the evaluation of ADPKD severity include demographic factors (age, gender, race, hypertension history, etc.), clinical factors (glomerular hyperfiltration, gross hematuria, cyst rupture, etc.), and some of laboratory factors (proteinuria and microalbuminuria, serum copeptin levels, serum biomarkers, etc.) (Schrier, 2014).

One of the key features of ADPKD is that the indexes of the kidney function remain in the normal range for several decades before a sharp decline (Grantham, 2006). This delayed index of kidney damage makes it difficult to diagnose and monitor disease progression at early stage, which is important for proper therapeutic intervention targeting early stages (Takiar, 2011). This feature of ADPKD emphasizes the significance of identifying markers for predicting early kidney damage and developing prediction methods based on available clinical and health information of patients.

A number of previous risk factor studies for ADPKD progression have mostly used linear models to test the correlation between different factors and disease progression represented by the slope of eGFR decrease (Cirillo, 2005). Recent studies have relied on mixed effect linear models in investigating the predictive roles of metabolic reprogramming, caffeine intake, serum galectin-3 level and other factors (Kim, 2019; McKenzie, 2018; Ozkurt, 2019). In most cases, age, race, and sex were treated as covariates. From a prediction point view, the overall predictive power of these models was low due to the non-linear nature of kidney function decline. Traditional paradigm of GFR progression was assumed to be steady over time, while many patients with chronic kidney disease have a nonlinear GFR trajectory or a prolonged period of non-progression (Li, 2012). Other methods were used to predict the progression of ADPKD towards ESRD with better results, including Bayesian smoothing techniques under posterior probability, and stochastic simulation models under assumption of fixed-time increment designed based on disease progression equations (Dulhare, 2016; Waezizadeh, 2018).

With the evolution in artificial intelligence and digital technology and the collection of massive amounts of health record data, machine learning has attracted the attention of healthcare researchers in solving core information processing and decision-making problems across a health system (Panch, 2018). For example, many machine learning classification and regression approaches, supervised and unsupervised, have been utilized in the field of diabetes studies from pathology research, laboratory diagnosis, to disease prediction, validation and feature selection (Alloghani, 2019; Rashidi, 2019; Maniruzzaman, 2018). Machine learning algorithms were originally designed as classifiers (Michie, 1994). Representative

methods include linear discriminant analysis, quadratic discriminant analysis, naïve Bayes, Gaussian process classification, support vector machine, artificial neural network, Adaboost, decision tree, and random forest. Several of these methods have better classification accuracy in biomedical studies than logistic regression, especially when data is sufficiently large and clean (Churpek, 2016). The machine learning methods are more flexible and sensitive to non-linearity and perform better than ordinary least squares (OLS). They have been developed for binary outcome and continuous outcome. Statistical test can be applied to determine significant differences between two machine learning algorithms (Segal, 2004; Trawiński, 2012; Cui, 2018).

This study was designed to apply machine learning approaches to predict the kidney function using both baseline and longitudinal data of patients in an ADPKD study cohort, US Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP). CRISP is a longitudinal study with average follow up over 7 years (Chapman, 2003). Our goal is to predict the eGFR value and its decrease in the next clinic visit based on all the prior existing data for the patient. Data used for prediction include demographics, biomarkers, and imaging information. Comparisons of prediction power were made among different machine learning methods in comparison with linear regression. We also examined the prediction performance in patient subgroups with different kidney function levels. Finally, important factors were identified that contributing to the prediction of eGFR and its decline.

Method

2.1 Data and Preprocessing

This study used CRISP data which is a unique longitudinal observational cohort of 242 ADPKD patients that began in 2000. The broad objectives of CRISP are to define the natural history of the disease and to discover prognostic biomarkers in early disease that can accurately predict long-term renal outcomes. The goals of CRISP are to continually follow-up the CRISP cohort so as to refine models of chronic kidney disease progression and strengthen the association between total kidney volume and renal outcomes, to validate disease models using existing and additional follow-up data from the HALT PKD study, and to incorporate powerful emerging imaging, genetic and biochemical biomarkers to improve the accuracy of prognostication for individual patients.

The CRISP data used for this study contained missing records which required filtering and imputation for next step model building. We removed some variables from the predictors that missed more than 25% of the data (3 variables: "Serum LDL cholesterol", " Urine Protein Concentration", " Urine Protein Excretion"). Any subject with more than 10 missing observations were deleted (11 subjects). The missing data in the rest of cohort were filled using imputation techniques.

MICE (Multivariate Imputation via Chained Equations) is one of the most principle methods for data imputation in R (Azur, 2011). It assumes that missing data are missed at random, which fits the pattern of the remaining missing data of CRISP. The probability of missing depends only on observed value and can be predicted using available observation. It imputes data on a variable-by-variable basis by specifying an imputation model per variable that contains missing records. Missing values are replaced by simulated draws from the posterior predictive distribution known as proper imputation. By default, linear regression is used to predict continuous missing values, and logistic regression is used for categorical missing values. The process is repeated over all variables with missing values in turn. Multiple imputation generated multiple complete datasets by filling in the missing value multiple times. MICE is very flexible to a broad range of missing types and more accurate in estimation with small standard error.

2.2 Machine Learning Methods

We used some machine learning methods implemented in the R package SuperLearner (SL) (Polley, 2019). This package has a large number of standard machine learning methods with the capacity of creating optimal weighted average of multiple models. It uses cross-validation to estimate the performance of multiple machine learning models, or the same model with different settings. The optimal weighted average approach has been proven to be asymptotically as accurate as the best possible prediction algorithm that is tested. This study is focused on four methods implemented in SuperLearner, Lasso regression, Random Forest, Support Vector Machine, and linear regression.

2.2.1 Simple linear regression

Simple linear regression is a supervised machine learning algorithm that can be performed by the SuperLearner library SL.lm. Linear regression performs the task to predict a dependent variable value based on given independent variables by achieving the best-fit regression line. The model aims to predict dependent value such that the error difference between predicted value and true value is minimum. The cost function of linear regression is the squared error between predicted value and the true value. It is reduced by starting with random values of coefficient and iteratively updating the values to reach the minimum cost.

2.2.2 Lasso Regression

Lasso (least absolute shrinkage and selection operator) regression is a penalized regression using elastic net performed by the SuperLearner library SL.glmnet. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients (Tibshirani, 1996). This type of regularization can result in sparse models with few coefficients: some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. λ is a tuning parameter that controls the strength of L1 penalty.

For $\lambda = 0$ it reduced to simple linear regression. Default setting of SuperLearner package is 10-fold cross validation and 100 checking for λ values.

2.2.3 Random Forest

Random Forest is a classification or regression algorithm consisting of many decisions trees which can be performed by the SuperLearner library SL.randomForest. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees. Its prediction by the ensemble of the trees is more accurate than that of any individual tree (Liaw, 2002). It draws bootstrap samples from the training data and chooses the best split among all the predictors for each node. It finally grows an unpruned classification or regression tree from the ensemble of the bootstrapped trees. The default size of terminal node is 5 for regression in SuperLearner package.

2.2.4 Support Vector Machine

Support Vector Machine (SVM) constructs a hyperplane or set of hyperplanes in a high dimensional space, which can be used for classification and regression performed by the SuperLearner library SL.svm. It is achieved by the best separation of hyperplane that has the largest distance to the nearest training data point of any class. For a finite-dimensional space the sets to discriminate are not linearly separable, so it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space to make the separation easier. The mappings used by SVM schemes are designed to ensure that dot products of pairs of input data vectors can be computed easily for the variables in original space, by defining them in terms of a kernel function. The sum of kernels can be used to measure the relative nearness of each test point to the data points originating in one or the other of the sets to be discriminated (Cortes, 1995).

2.3 Model Construction and Study Design

2.3.1 Predictors

We used age, education, gender, race from demographic dataset, 60 variables from biomarker dataset, total kidney volume (TKV) from imaging dataset and eGFR value of previous visit(s) to construct models for predicting eGFR value at visit or the eGFR drop (ΔeGFR) between two consequent visits.

2.3.2 Model construction

The prediction is achieved by the method of leave-one-out cross validation: for each observation in the data, take the other observations as training dataset to build up a machine learning model for predicting the outcome value of the test data which is the omitted observation.

For model assessment and comparison, MSE and R^2 are calculated for each model (y represent observed value, $\hat{y}$ predicted value):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i), \quad R^2 = corr(y_{Observed}, \hat{y}_{Predicted}\})^2$$

The model construction for prediction can be expressed in the flowchart in Figure 1.

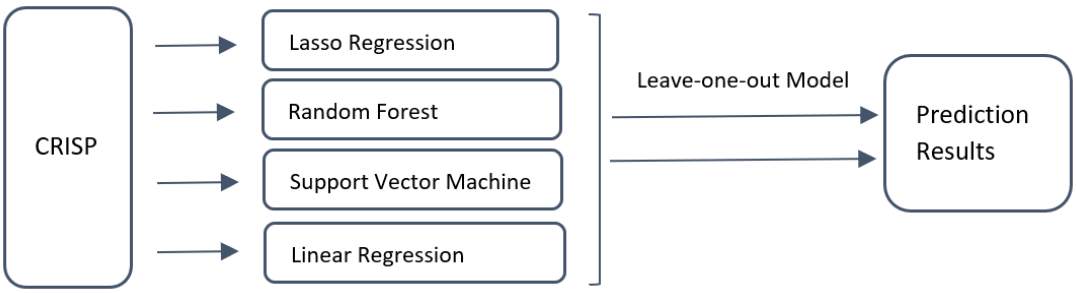


Figure 1. Model Construction Flow Chart

2.3.3 Study Design

This study is consisted of four major parts of analysis: single-year prediction, patients subgrouping, multiple-year prediction and important variables identification.

2.3.3.1 Prediction of eGFR value & eGFR change using one year of data

Predict based on the previous year data. The outcomes to be predicted are eGFR value or eGFR drop of Year 2. The predictors are Year 1 clinical records of demographic, biomarker variables, total kidney volume (TKV) from imaging, and eGFR value. The models structure can be simplified as following expressions:

$$eGFR_2 \sim Demo + Biomarker_1 + TKV_1 + eGFR_1 \quad (1)$$

$$\Delta eGFR_2 \sim Demo + Biomarker_1 + TKV_1 + eGFR_1 \quad (2)$$

Here $\Delta eGFR_2 = eGFR_2 - eGFR_1$.

Predict into the future. Here we chose to use Year 2 information to predict Year 6 eGFR value because the eGFR records after Year 6 have a higher missing rate, and the span of four years is considered as a long period for prediction. The models structure can be simplified as following expressions:

$$eGFR_6 \sim Demo + Biomarker_2 + TKV_2 + eGFR_2 \quad (3)$$

Evaluate the contribution of the previous eGFR drop. The eGFR drop of Year 0 to Year 1 was added as predictive variable in the one-year prediction of Year 2 eGFR value using Year 1 information and compared with model 1. The model structure can be simplified as following expression:

$$eGFR_2 \sim Demo + Biomarker_1 + TKV_1 + eGFR_1 + \Delta eGFR_1 \quad (4)$$

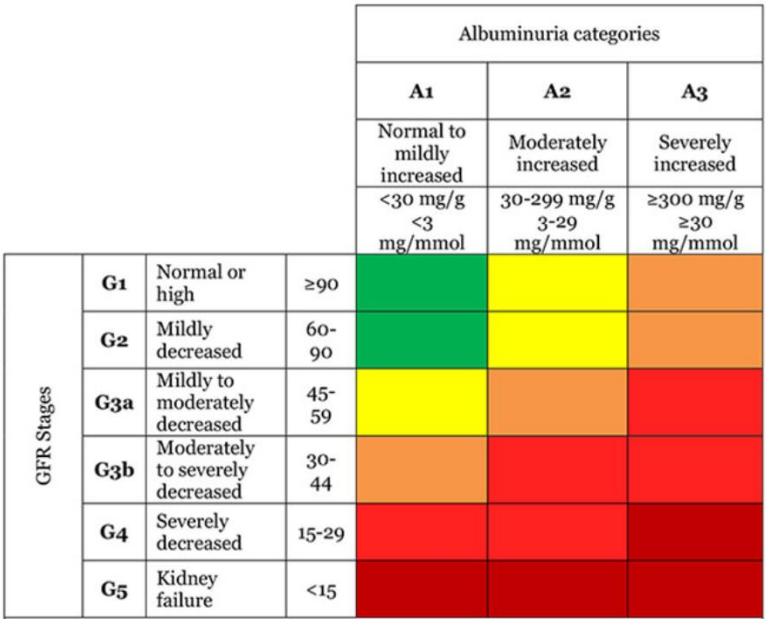
Here $\Delta eGFR_1 = eGFR_1 - eGFR_0$.

2.3.3.2 Patients subgrouping according to kidney function level

Generally, the normal eGFR value is more than 90 mL/min/1.73m2 in adults. eGFR declines with age, even in people without kidney disease. Other testing should be used to affirm the result of eGFR examination. Blood or protein in the urine can be an early sign of kidney disease. People with a high

amount of albumin in their urine (albuminuria) are at an increased risk of having chronic kidney disease progress to kidney failure (Figure 1).

Subjects in the CRISP cohort started initial clinical visit at different kidney function stages. Some were at early detection of kidney deterioration and some were near kidney failure. Therefore in this study, they were divided into two subgroups according to their kidney function stage: non-normal group with eGFR<90 and normal group with eGFR>90. The two groups were applied to Lasso regression models of predicting Year 2 eGFR value or eGFR change to compare the performance of subgrouping, model structure is the same as model (1) and (2).



* Colors represent the risk of progression, morbidity and mortality from best to worst. Green: low risk. Yellow: moderately increased risk. Orange: high risk. Red: very high risk. Deep red: highest risk.[13]

Figure 2. Kidney Function Stages by GFR and Albuminuria

2.3.3.3 Prediction of eGFR value using multiple years of information

More years of clinical information was used for the prediction of eGFR value and compared with using only one year of information. This section was focused on predicting Year 3 kidney function by using only Year 2 information, using Year 1 and Year 2 information, and using Year 0 to Year 2 information.

Predictors include demographic information, 1-3 years of biomarker information, 1-3 years of total kidney volume record(s) and 1-3 years of eGFR record(s). The model structure can be expressed as:

$$eGFR_3 \sim Demo + Biomarker_2 + TKV_2 + eGFR_2 \quad (5)$$

$$eGFR_3 \sim Demo + Biomarker_{1,2} + TKV_{1,2} + eGFR_{1,2} \quad (6)$$

$$eGFR_3 \sim Demo + Biomarker_{0,1,2} + TKV_{0,1,2} + eGFR_{0,1,2} \quad (7)$$

2.3.3.4 Important variables identification

In order to test the necessity of including two expensive factors in prediction model: PKD genotype and total kidney volume which are expensive to obtain in clinics, prediction models of Year 2 eGFR value were computed using Year 1 information without these factors. It was compared between models using all variables and models excluding genotype only, excluding total kidney volume only and excluding both two factors.

Health information including mass index (BMI), mean article pressure (map), pain in the left or right kidney (pain), mean of seated and standing systolic blood pressure (systol) and urinary tract infection (uti) from the health dataset were considered as predictors in the model predicting Year 3 eGFR. Multiple-year prediction was performed by using only Year 2 information, using Year 1 and Year 2 information, and using Year 0 to Year 2 information. The model structure can be expressed as:

$$eGFR_3 \sim Demo + Biomarker_2 + Health_2 + eGFR_2 \quad (8)$$

$$eGFR_3 \sim Demo + Biomarker_{1,2} + Health_{1,2} + eGFR_{1,2} \quad (9)$$

$$eGFR_3 \sim Demo + Biomarker_{0,1,2} + Health_{0,1,2} + eGFR_{0,1,2} \quad (10)$$

Important variables that affect the prediction were identified by non-zero coefficients from Lasso regression model for Year 2 eGFR value and eGFR change using Year 1 information. Data of Year 1

predictors was standardized (subtracted the mean and divided by the standard diviation) using R Scale function to make the Lasso coefficients values comparable. 10-fold cross validation was performed for the cohort of 242 subjects. The SuperLearner package generated 100 models by default for each fold. The median of beta coefficients from 100 models was selected to compare the significance of variables.

Result

The CRIPS dataset contains 242 subjects with 8 to 11 years of longitudinal observations for each subject. The cohort is consisted of 97 males and 145 females, and majority of them are Caucasian (209). Their average age at baseline is 32.4, average years of education is 14.7 (Table 1).

Categorical Variable	Mean (SD)
Age	32.37 (8.89)
Year of Education	14.63 (3.26)
Continuous Variable	Number (Percentage)
Gender = Male	97 (39.8)
Female	145 (60.2)
Race = Caucasian	209 (86.7)
African	28 (11.6)
Hispanic	2 (0.8)
Asian	2 (0.8)

Table 1. Basic Statistical Summary of Demographic Information

In this cohort, there is substantial missing follow up. Demographic data contains sporadic missing records (<1%) for education and race, and complete for other variables. The eGFR records are complete for Year 0 to Year 6. Follow up of eGFR was not conducted on most patients for Year 7 and Year 8 but continued in later years (Year 9 to Year 11). The missing rate for biomarker variables is low (6.5%) in the first few years (Year 0 to Year 3) while high (above 90%) for the following up years. This study is mostly focused on the prediction of kidney function in the early years of follow up (Year 2, Year 3 and Year 6). After

removing 11 subjects with extensive missing of biomarker records, 231 subjects and 58 predictor variables were used to predict eGFR value.

3.1 Patient eGFR can be predicted well with previous year's information

To assess the prediction power of previous year's data on the current year eGFR, we used four different machine learning methods (Lasso regression, Random Forest, Support Vector Machine and Linear Regression models) to predict the eGFR values of Year 2 using Year 1 information. Leave-one-out cross validation was used to compare the machine learning methods and predict the outcome eGFR value or eGFR drop for all subjects. The prediction power comparison was based on mean square error (MSE) and R squared (R^2) calculated from the predicted and observed Year 2 eGFR value. The results showed that high prediction power for eGFR, more than 64% in R^2 for all methods (Table 2). Among the four machine learning methods, Lasso is most accurate (R^2=0.78) and random forest performs slightly lower. Both these methods are better than the linear regression. The MSE values are consistent with the R^2 results.

We also examined the prediction of $\Delta eGFR_2$ (yearly drop of eGFR value of Year 2) similarly. However, the prediction power is lower than Year 2 eGFR, with R^2 value near zero for all methods.

		Machine Learning Models			
Outcome		Lasso	Random Forest	SVM	Linear Regression
Year 2 eGFR	MSE	130.043	139.705	216.827	173.259
	R^2	0.782	0.777	0.647	0.718
eGFR Delta	MSE	184.095	181.832	194.946	277.159
	R^2	0.024	0.039	0.002	0.003

Table 2. Year1 to Year2 Prediction of eGFR value and eGFR Drop

3.2 Prediction accuracy decreases for more distant future time-points

To assess the extent of prediction power projecting into the future, we compared the prediction of eGFR value for one year later and four years later. The one-year prediction is the same as Table 2 predicting

Year 2 eGFR value using Year 1 information. The four-year prediction is conducted to predict Year 6 eGFR value using Year 2 information with the four machine learning methods and leave-one-out cross validation.

For all machine learning methods, the MSE of four-year prediction (Year 2 to Year 6) is about 2.3 to 6.5 times larger than that of one-year (Year 1 to Year 2) prediction, correspondingly the R^2 of four-year prediction is much smaller than that of one-year prediction (Table 3). The predicted value for Year 6 eGFR using Year 2 data is less accurate than that for Year 2 eGFR using Year 1 data (Figure 3). The comparison of MSE and R^2 between two prediction of different extent showed that the prediction accuracy drops dramatically when predicting into the further future as expected. For Lasso regression, random forest and support vector machine methods, the R^2 dropped about 40% for four-year prediction compared to one-year prediction, and for linear regression the R^2 dropped about 70%. This indicates the linear regression method is less stable for prediction of eGFR projecting into the far future.

		Machine Learning Models			
	Model	Lasso	Random Forest	SVM	Linear Regression
MSE	Yr1 -> Yr2	130.043	139.705	216.827	173.259
	Yr2 -> Yr6	481.012	446.897	509.055	1140.701
R^2	Yr1 -> Yr2	0.782	0.777	0.647	0.718
	Yr2 -> Yr6	0.425	0.470	0.391	0.218

Table 3. One-Year and Four-Year Machine Learning Prediction of eGFR Value

3.3 Previous year eGFR change contributes to improve the prediction power

To find potential improvement for the prediction power of one-year prediction, the change of previous year's eGFR was examined in its contribution to predicting eGFR values. $\Delta eGFR_1$ (yearly drop of eGFR value of Year 1) was added as a predictor variable into the machine learning models for predicting Year 2 eGFR using Year1 information, and compared with the models not using $\Delta eGFR_1$.

The resulted R^2 values improved for all methods, where the improvement is relatively large for Lasso regression (3.6%), support vector machine (2.2%) and, but much smaller for random forest (1.3%) and linear regression (1.7%). The MSE results consistently increased when including Year 0 to Year 1 eGFR drop as a predictor (Table 4). The comparison suggests previous year's eGFR change contributed to improving the prediction power of eGFR value.

Model		Machine Learning Models			
		Lasso	Random Forest	SVM	Linear Regression
Without $\Delta eGFR_1$	MSE	130.043	139.705	216.827	173.259
	R^2	0.782	0.777	0.647	0.718
With $\Delta eGFR_1$	MSE	113.104	135.339	209.698	168.422
	R^2	0.810	0.787	0.661	0.730

Table 4. Comparison between With and Without ΔeGFR in Year1->Year2 eGFR Prediction Model

3.4 Subgrouping of patients lead to separated prediction performance

In order to look at the prediction performance for eGFR value and eGFR change on patients with different kidney function stage, the cohort was divided into two subgroups according to their Year 1 eGFR value (non-normal group with eGFR<90 and normal group with eGFR>90). The lower eGFR subgroup contains 109 subjects and the higher eGFR subgroup contains 122 subjects. The two subgroups were separately applied into the Lasso regression model of predicting Year 2 eGFR or $\Delta eGFR_2$ using Year 1 information.

For both subgroups of patients, the eGFR values can be better predicted than the eGFR change as expected. The difference between model R^2 is bigger for non-normal subgroup compared to normal subgroup (Table 5). The prediction for eGFR value is more accurate for non-normal subgroup (R^2=0.64) than normal subgroup (R^2=0.23). In contrast, the prediction for eGFR change is more accurate for normal subgroup (R^2=0.21) than non-normal subgroup (R^2=0.05). The prediction for eGFR change is highly biased and less accurate for the normal subgroup of patients (Figure 4). These results indicate that eGFR value performs best as the outcome and it can be better predicted for the non-normal patients.

Outcome		Subgroup of Patients	
		Non-normal Subgroup	Normal Subgroup
eGFR	R^2	0.640	0.233
ΔeGFR	R^2	0.046	0.210

Table 5. Comparison between Non-normal and Normal Subgroups for Year1 to Year2 Prediction of eGFR/ΔeGFR

3.5 Effect of adding additional year's historical information

To evaluate the contribution of additional historical data on prediction power, we conducted comparison between multiple-year prediction of Year 3 eGFR value using only Year 2 information, using Year 1 and Year 2 information, and using Year 0 to Year 2 information. For Lasso regression, random forest and support vector machine methods in predicting Year 3 eGFR, adding one year's historical information (Year 1) improved the prediction power. It increased R^2 (3%) and decreased MSE (10%) compared with using only Year 2 information (Table 6). However, this trend is different for linear regression method as the prediction power decreased when including Year 1 data. Adding another year of historical information (Year 0) continuously improved the prediction power for random forest with a smaller increasing rate of R^2 (1.7%). While for Lasso regression and support vector machine, adding more historical information actually reduced the prediction power represented by decreased R^2 and increased MSE.

Predictors		Machine Learning Models			
		Lasso	Random Forest	SVM	Linear Regression
Yr2	MSE	124.790	138.527	204.557	234.173
	R^2	0.782	0.768	0.657	0.643
Yr1 + Yr2	MSE	113.157	120.937	202.798	358.728
	R^2	0.802	0.795	0.670	0.538
Yr0 + Yr1 + Yr2	MSE	114.686	112.710	208.026	-
	R^2	0.799	0.809	0.662	-

***** The linear regression on using Year 0 to Year 2 was not performed due to collinearity.

Table 6. Prediction of Year3 eGFR Value Using 1-3 Years of Historical Information

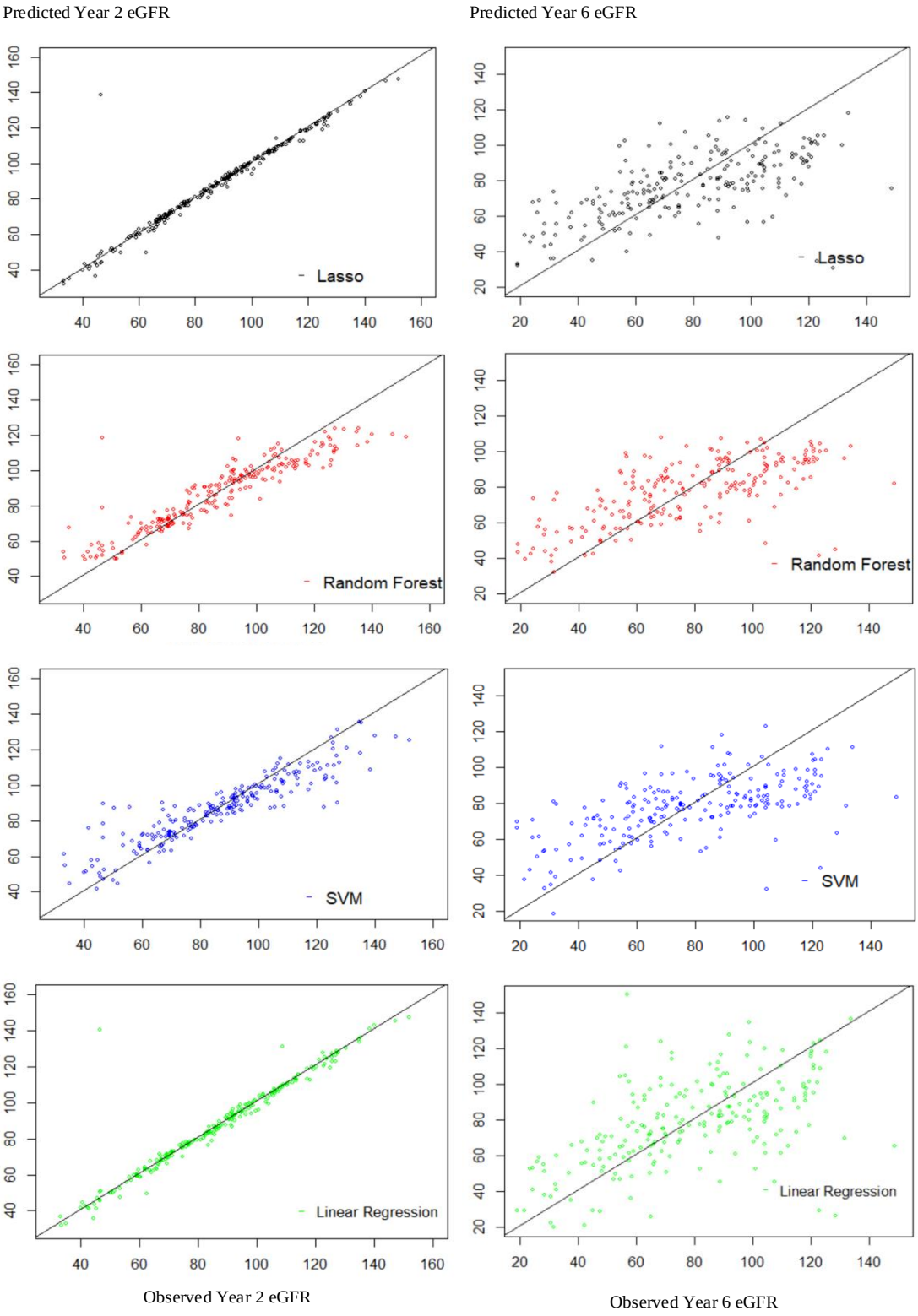


Figure 3. Observed vs Predicted eGFR Value for Single-Year Prediction. Plots on the left panel are one-year prediction of Year1 to Year2. Plots on the right panel are four-year prediction of Year2 to Year6. X-axis is observed eGFR value and y-axis is predicted eGFR value.

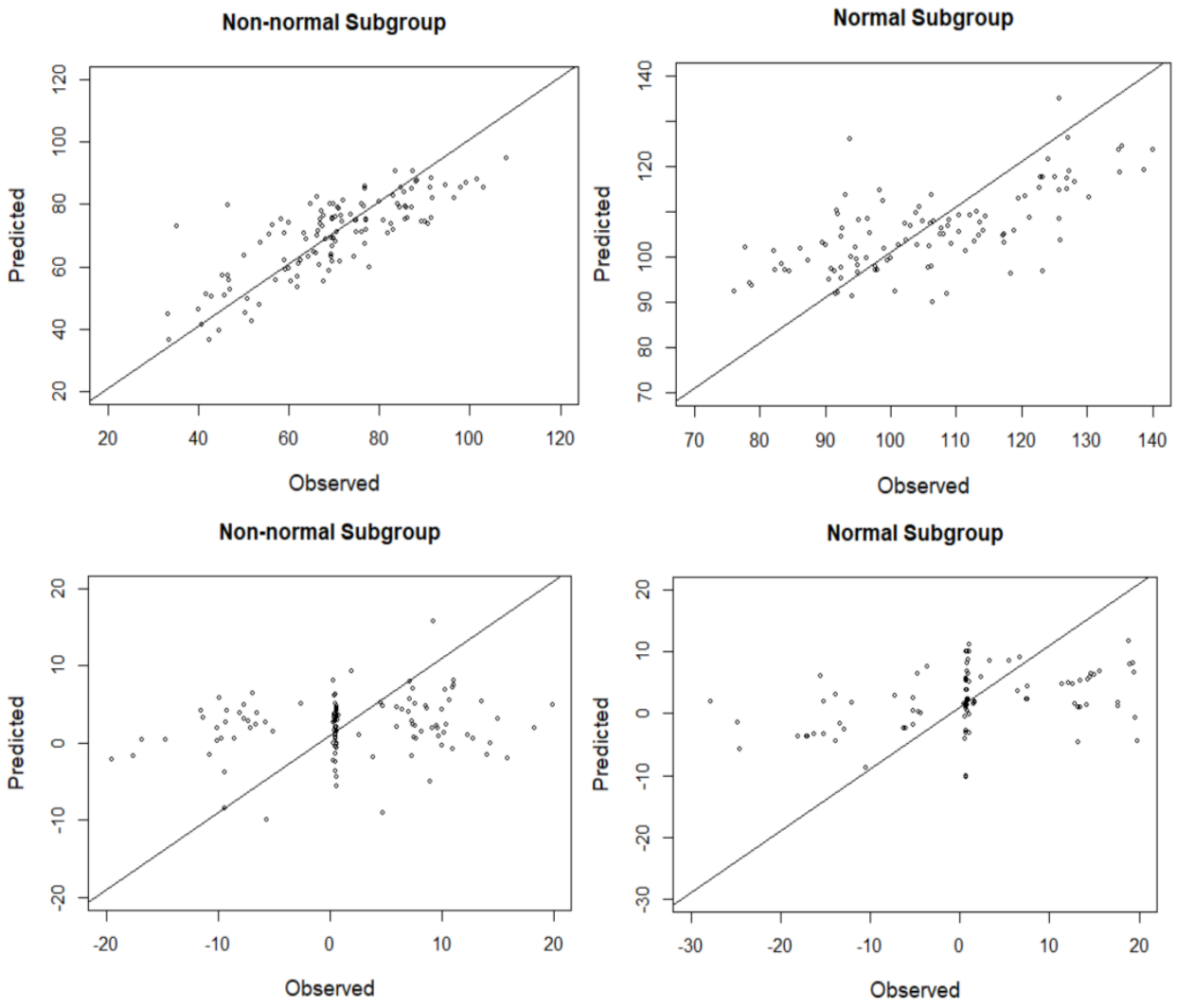


Figure 4. Observed vs. Predicted Year2 eGFR/ΔeGFR for Non-normal Subset with Lower Year1 eGFR and Normal Subset with Higher Year1 eGFR using Lasso Regression

3.6 Expensive genotypes and image predictors can be replaced

Genotypes of PKD genes and the total kidney volume estimated from imaging are expensive to generate in clinical settings. Their contribution to eGFR prediction might be replaced with other biomarkers routinely tested in clinics. We compared the performance between predicting Year 1 to Year 2 eGFR using all predictors available with models without PKD genotype, or total kidney volume, or both.

For Lasso regression, random forest and linear regression, the prediction power slightly increased by about 1.7% to 3.1% after dropping the expensive variables (Table 7). For support vector machine, the R^2

did not change dramatically regardless of the inclusion or exclusion of PKD genotype or total kidney volume or both. These results suggest that these two expensive variables can be replaced by biomarker variables routinely collected in clinics.

Machine Learning Model		Expensive Factors Usage			
		Include Both	Exclude Genotype	Exclude TKV	Exclude Both
Lasso	R^2	0.782	0.796	0.796	0.796
Random Forest	R^2	0.777	0.796	0.795	0.796
SVM	R^2	0.647	0.654	0.643	0.657
Linear Regression	R^2	0.718	0.748	0.740	0.750

Table 7. Comparison among Prediction Excluding PKD Genetype, Excluding TKV and Excluding Both and Including Both

3.7 Including Health information improves the prediction power slightly

The CRISP cohort collected health information on BMI, mean arterial pressure, pain, systolic blood pressure and urinary tract infection. We added these variables to the prediction models for single-year or multiple-year prediction to examine their contribution to the prediction power.

The results are shown in Table 8. Comparing to Table 6 resulted from the same models without these predictors, Table 8 shows slight increase in prediction power in general. For the Year 2 to Year 3 eGFR prediction, the R^2 increased 0.001 for Lasso regression and 0.008 for support vector machine when including health records, while it decreased for random forest and linear regression. For prediction models using both Year 1 and Year 2 data, adding health records actually slightly reduced the prediction power for Lasso regression and random forest models, while merely improved the power for support vector machine (0.008 increase in R^2) and linear regression (0.043 increase in R^2). For prediction using all three previous years of data, including health records performed similarly to models without health records.

Predictors		Machine Learning Models			
		Lasso	Random Forest	SVM	Linear Regression
Yr2	MSE	124.681	139.696	199.052	418.434
	R^2	0.783	0.765	0.665	0.486
Yr1 + Yr2	MSE	115.841	123.295	197.093	308.366
	R^2	0.799	0.790	0.678	0.581
Yr0 + Yr1 + Yr2	MSE	112.246	115.818	202.222	-
	R^2	0.805	0.803	0.671	-

Table 8. Prediction of Year3 eGFR Value Using 1-3 Years of Historical Information with Health Records

4 Important predictors for eGFR value and eGFR change

To identify variables that contribute to the prediction of eGFR values, we plotted the non-zero median coefficients of the predictors in Lasso regression for prediction of Year 2 eGFR using standardized Year 1 data (Figure 5). The largest contribution comes from the Year 1 eGFR value, which is consistent with the fact that eGFR changes slowly. Other smaller but consistent contributors are gender, previous year eGFR change, copeptin, PKD genotype, serum creatinine concentration, and urine potassium concentration. There are 24 other predictors that contribute to the prediction but to a much less degree. About half of the non-zero contributors have positive coefficients and half of them have negative coefficients.

Similarly, we plotted the non-zero coefficients of predictors for predicting $\Delta eGFR_2$ using standardized Year 1 data. Given the low prediction power for eGFR drop, the coefficients are much smaller. The top ranked contributors include gender, serum creatinine concentration, Year 1 eGFR value, previous year eGFR change, gene type, urine potassium concentration, urine magnesium concentration, and urine sodium excretion. There are 23 other predictors that contribute to the prediction but to a much less degree. About 2/3 of the non-zero contributors have positive coefficients and 1/3 of them have negative coefficients.

Discussion

This study performed prediction of eGFR value and eGFR change for the CRISP cohort using different historical information and different types of clinical variables. Four machine learning methods were evaluated, Lasso regression, random forest, support vector machine and linear regression. We found that predicting eGFR values using previous year's information is much more powerful than prediction on the eGFR yearly change. The prediction performed better for patients with abnormal kidney function than patients with normal kidney functions in subgroup analysis. The prediction power for eGFR values decreased when projecting into the distant future. Including additional predictors, such as health variables, and previous year eGFR change helped to a small improvement. The expensive predictors of PKD genotype and total kidney volume can be replaced by other biomarker variables routinely collected in clinics without affecting the prediction power of machine learning models.

One result that caught our attention is the decreased R^2 when subgroupping the patients according to their kidney function stages. Building models separately for patients in normal and non-normal groups did not lead to improved prediction power for both groups compared with the complete cohort, while the machine learning models had distinguished performance on different groups. Another result is outstanding point in Figure 3 Year 1 to Year 2 prediction, that Lasso regression, random forset and linear regression pointed out a suspect outlier. While SVM predicted well on that point, suggesting stable performance of SVM on extreme observations. A limitation of this study is the small sample size of CRISP cohort, which made the prediction hard to be applied with several machine learning approaches, such as Neurol Network that requires a large sample size to build up the models.

In future studies, more tests should be conducted on using predictive variables from imaging results and medication records and evaluating their contribution in improving the machine learning model performance. Further works can be focused on the extent of historical information usage in predicting eGFR values. This would provide support for the threshold of clinical follow up years and data collection period, balancing between cost effectiveness maximization of prediction power.

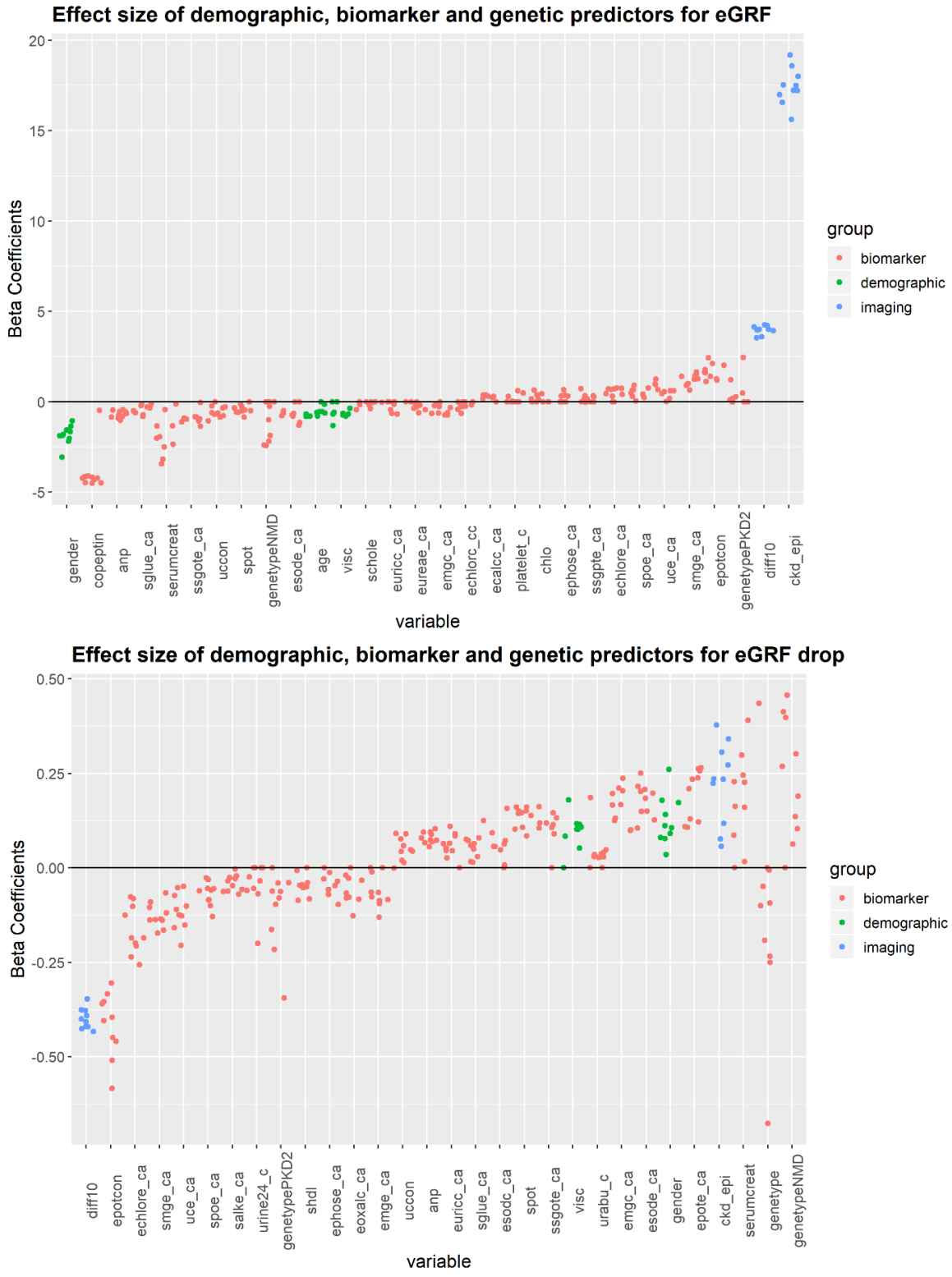


Figure 5. Important variables contributing variables to the prediction of eGFR values and eGFR changes in Biomarkers, Demographic and Imaging. (Appendix)

Reference:

1. Alloghani, M., Aljaaf, A., Hussain, A., Baker, T., Mustafina, J., Al-Jumeily, D., & Khalaf, M. (2019). Implementation of machine learning algorithms to create diabetic patient re-admission profiles. BMC Medical Informatics and Decision Making, 19(9), 253.

2. Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research, 20(1), 40-49.

3. Bajwa, Z. H., Sial, K. A., Malik, A. B., & Steinman, T. I. (2004). Pain patterns in patients with polycystic kidney disease. Kidney international, 66(4), 1561-1569.

4. Chapman, A. B., Guay-Woodford, L. M., Grantham, J. J., Torres, V. E., Bae, K. T., Baumgarten, D. A., ... & Brummer, M. E. (2003). Renal structure in early autosomal-dominant polycystic kidney disease (ADPKD): The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) cohort. Kidney international, 64(3), 1035-1045.

5. Chebib, F. T., & Torres, V. E. (2016). Autosomal dominant polycystic kidney disease: core curriculum 2016. American Journal of Kidney Diseases, 67(5), 792-810.

6. Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W., & Edelson, D. P. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Critical care medicine, 44(2), 368.

7. Cirillo, M., Anastasio, P., & De Santo, N. G. (2005). Relationship of gender, age, and body mass index to errors in predicted kidney function. Nephrology Dialysis Transplantation, 20(9), 1791-1798.

8. Collins, A. J. FR (2012). US renal data system 2011 Annual data report. Am J Kidney Disease, 59(1 SUPPL 1), 6386.

9. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

10. Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. Neuroimage, 178, 622-637.

11. Dulhare, U. N., & Ayesha, M. (2016, December). Extraction of action rules for chronic kidney disease using Naïve bayes classifier. In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-5). IEEE.

12. Eisenberger, T., Decker, C., Hiersche, M., Hamann, R. C., Decker, E., Neuber, S., ... & Toenshoff, B. (2015). An efficient and comprehensive strategy for genetic diagnostics of polycystic kidney disease. PloS one, 10(2).

13. Estimated Glomerular Filtration Rate (eGFR). (2019, January 15). Retrieved from https://www.kidney.org/atoz/content/gfr

14. Grantham, J. J., & Torres, V. E. (2016). The importance of total kidney volume in evaluating progression of polycystic kidney disease. Nature Reviews Nephrology, 12(11), 667.

15. Grantham, J. J., Torres, V. E., Chapman, A. B., Guay-Woodford, L. M., Bae, K. T., King Jr, B. F., ... & Klahr, S. (2006). Volume progression in polycystic kidney disease. New England Journal of Medicine, 354(20), 2122-2130.

16. Helal, I. (2013). Autosomal dominant polycystic kidney disease: new insights into treatment. Saudi Journal of Kidney Diseases and Transplantation, 24(2), 230.

17. Higashihara, E., Horie, S., Muto, S., Mochizuki, T., Nishio, S., & Nutahara, K. (2012). Renal disease progression in autosomal dominant polycystic kidney disease. Clinical and experimental nephrology, 16(4), 622-628.

18. Kim, K., Trott, J. F., Gao, G., Chapman, A., & Weiss, R. H. (2019). Plasma metabolites and lipids associate with kidney function and kidney volume in hypertensive ADPKD patients early in the disease course. BMC nephrology, 20(1), 66.

19. Li, L., Astor, B. C., Lewis, J., Hu, B., Appel, L. J., Lipkowitz, M. S., ... & Greene, T. H. (2012). Longitudinal progression trajectory of GFR among patients with CKD. American journal of kidney diseases, 59(4), 504-512.

20. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

21. Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of medical systems, 42(5), 92.

22. McKenzie, K. A., El Ters, M., Torres, V. E., Harris, P. C., Chapman, A. B., Mrug, M., ... & Alan, S. L. (2018). Relationship between caffeine intake and autosomal dominant polycystic kidney disease progression: a retrospective analysis using the CRISP cohort. BMC nephrology, 19(1), 378.

23. Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning. Neural and Statistical Classification, 13(1994), 1-298.

24. Moyer, J. H., Lee-Tischler, M. J., Kwon, H. Y., Schrick, J. J., Avner, E. D., Sweeney, W. E., ... & Woychik, R. P. (1994). Candidate gene associated with a mutation causing recessive polycystic kidney disease in mice. Science, 264(5163), 1329-1333.

25. Ozkurt, S., Dogan, I., Ozcan, O., Fidan, N., Bozaci, I., Yilmaz, B., & Bilgin, M. (2019). Correlation of serum galectin-3 level with renal volume and function in adult polycystic kidney disease. International urology and nephrology, 51(7), 1191-1197.

26. Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. Journal of global health, 8(2).

27. Polley, E., LeDell, E., Kennedy, C., Lendle, S., & van der Laan, M. (2019). Package 'SuperLearner'.

28. Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., & Green, R. (2019). Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. Academic pathology, 6, 2374289519873088.

29. Schrier, R. W., Brosnahan, G., Cadnapaphornchai, M. A., Chonchol, M., Friend, K., Gitomer, B., & Rossetti, S. (2014). Predictors of autosomal dominant polycystic kidney disease progression. Journal of the American Society of Nephrology, 25(11), 2399-2418.

30. Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

31. Stevens, L. A., Coresh, J., Greene, T., & Levey, A. S. (2006). Assessing kidney function—measured and estimated glomerular filtration rate. New England Journal of Medicine, 354(23), 2473-2483.

32. Takiar, V., & Caplan, M. J. (2011). Polycystic kidney disease: pathogenesis and potential therapies. Biochimica ET Biophysica Acta (BBA)-Molecular Basis of Disease, 1812(10), 1337-1343.

33. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

34. Trawiński, B., Smętek, M., Telec, Z., & Lasota, T. (2012). Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. International Journal of Applied Mathematics and Computer Science, 22(4), 867-881.

35. Waezizadeh, T., Mehrpooya, A., Rezaeizadeh, M., & Yarahmadian, S. (2018). Mathematical models for the effects of hypertension and stress on kidney and their uncertainty. Mathematical biosciences, 305, 77-95.

Appendix. Variable Explanation

Dataset	Variable*	Obs	Unique	Mean	Min	Max	Label**
Demographic	visc	2375	1374	5.8333	0	13.2046	Number of years since CRISP I baseline visit date
Demographic	age	2375	2160	38.40	14.78	58.07	Age at visit
Demographic	educ	2507	21	15.09	7.00	28.00	Total number years of education
Demographic	gender	2511	2	1.60	1.00	2.00	Gender
Demographic	race4	2511	4	1.14	1.00	4.00	Race 1=Caucasian 2=African American 3 = Hispanic 4=Asian
Imaging	tkv	1383	1337	1366.93	276.16	8711	Total Kidney Volume (ml)
Imaging	ckd_epi	.	.	.	.	.	eGFR calculated based on CKD_epi
Biomarker	albe_ca	820	671	45.95	0.76	736.00	Urine: Albumin Excretion (mg/24hr)
Biomarker	anp	1785	224	47.78	14.88	153.74	proANP at baseline (pmol/l)
Biomarker	chlo	1480	21	104.10	93.00	113.00	Chloride
Biomarker	co2	1480	23	25.58	8.00	36.00	CO2
Biomarker	copeptin	1826	196	4.74	0.52	143.00	Copeptin at baseline (pmol/l)
Biomarker	ecalcc_ca	823	585	7.87	0.08	158.00	Urine: Calcium Concentration (mg/dL)
Biomarker	ecalce_ca	818	590	164.58	3.00	2291.00	Urine: Calcium Excretion (mg/24h)
Biomarker	echlorc_cc	863	409	57.31	1.57	213.00	Urine: Chloride Concentration (mEq/dL)
Biomarker	echlore_ca	859	648	179.05	3.35	5520.64	Urine: Chloride Excretion (mEq/24h)
Biomarker	ecitrc_ca	870	740	66.10	0.32	756.00	Urine: Citrate Concentration (mg/dL) (c) in CRISP I
Biomarker	ecitre_ca	866	740	460.09	0.83	2984.00	Urine: Citrate Excretion (mg/24h)
Biomarker	emgc_ca	814	550	5.29	0.19	136.31	Urine: Magnesium Concentration (mg/dL)
Biomarker	emge_ca	810	532	300.06	4.79	4866.18	Urine: Magnesium Excretion (mg/24h)
Biomarker	eoxalc_ca	865	443	4.95	0.05	66.00	Urine: Oxalate Concentration (mg/dL)
Biomarker	eoxale_ca	861	445	31.22	2.00	237.00	Urine: Oxalate Excretion (mg/24h)
Biomarker	ephosc_ca	902	601	42.92	1.60	713.00	Urine: Phosphorus Concentration (mg/dL)
Biomarker	ephose_ca	898	768	961.00	39.70	19429.00	Urine: Phosphorus Excretion (mg/24h)
Biomarker	epotcon	897	344	25.38	1.88	115.00	Urine Potassium concentration

Biomarker	epote_ca	892	537	57.47	1.12	176.00	Urine: Potassium Excretion (mEq/24h)
Biomarker	esodc_ca	899	501	174.51	14.06	515.87	Urine: Sodium Concentration (mg/dL)
Biomarker	esode_ca	900	703	4339.78	0.00	13692.84	Urine: Sodium Excretion (mg/24h)
Biomarker	esode_cc	894	698	189.71	1.00	595.60	Urine: Sodium Excretion (mEq/24h)
Biomarker	eureac_ca	897	774	403.35	0.14	1402.27	Urea Concentration (mg/dL)
Biomarker	eureae_ca	868	589	9508.39	2033.36	40857.20	Urine: Urea Nitrogen Excretion (mg/24h)
Biomarker	euricc_ca	897	524	27.23	0.86	191.01	Urine: Uric acid Conc (mg/dL)
Biomarker	eurice_ca	892	747	611.46	22.30	4294.00	Urine: Uric acid Excretion (mg/24h)
Biomarker	genetype	1873	3	.	.	.	PKD Gene Mutation
Biomarker	hemoglob	911	72	13.34	8.20	17.00	Hemoglobin
Biomarker	hemotocrit	911	169	38.84	28.60	48.50	Hematocrit
Biomarker	il18	456	20	1.00	0.00	9.60	Interleukin-18 at baseline
Biomarker	lptrie_ca	879	249	124.35	23.00	814.00	Serum: Triglyceride (mg/dL)
Biomarker	ngal	468	40	34.31	0.04	476.21	Neutrophil gelatinase-associated lipocalin at baseline
Biomarker	platelet_c	910	321	221.96	0.14	540.00	Platelet (K/uL)
Biomarker	salke_ca	915	157	71.75	24.00	530.00	Serum: Alk Phos (U/L)
Biomarker	sbilire_ca	909	23	0.66	0.05	2.30	Serum: Bilirubin (mg/dL)
Biomarker	sbune_ca	913	32	14.72	4.00	40.00	Serum: BUN (mg/dL)
Biomarker	scalc	916	24	9.14	8.00	10.30	Serum calcium
Biomarker	schole	1440	173	173.07	68.00	346.00	Serum total cholesterol
Biomarker	serumcreat	1755	214	1.17	0.00	8.60	Serum creatinine concentration
Biomarker	sglue_ca	915	79	88.16	30.00	166.00	Serum: Glucose (mg/dL)
Biomarker	shdl	1439	82	46.37	14.00	152.00	Serum HDL cholesterol
Biomarker	sldl	1280	177	102.00	16.00	306.80	Serum LDL cholesterol
Biomarker	smge_ca	895	25	1.93	1.22	2.68	Serum: Magnesium (mg/dL)
Biomarker	sod	1483	21	137.84	125.00	145.00	Sodium
Biomarker	spoe_ca	908	38	3.64	1.40	5.60	Serum:PO4 (mg/dL)
Biomarker	spot	916	36	3.95	2.50	34.00	Serum potassium
Biomarker	ssgote_ca	914	51	23.35	5.00	328.00	Serum: SGOT(AST) (U/L)
Biomarker	ssgpte_ca	906	67	22.03	6.00	360.00	Serum: SGPT(ALT) (U/L)
Biomarker	surice_ca	911	84	5.19	0.90	14.00	Serum: Uric Acid (mg/dL)
Biomarker	trunc_grp	1738	2	.	.	.	Truncated or non-truncated gene mutation
Biomarker	uccon	901	434	72.14	7.66	422.00	Urine Creatinine concentration

Biomarker	uce_ca	896	776	1578.54	90.00	9528.80	Urine: Creatinine Excretion (mg/24)
Biomarker	uprotc_ca	601	104	10.79	0.15	274.00	Urine: Protein concentration (mg/dL)
Biomarker	uprote_ca	596	463	228.60	6.00	5181.30	Urine: Protein Excretion (mg/24hr)
Biomarker	urabu_c	1351	627	740195.30	0.00	1000000000.00	Unit corrected of urabu
Biomarker	urine24_c	903	786	2634.30	449.00	7150.00	24-hour Urine Volume
Biomarker	urine_mcp	1793	225	603.61	19.80	4517.80	Urine mcp at baseline
Biomarker	wbc_c	911	204	5.77	0.00	15.00	White Blood Cell (K/uL)

* Variable is variable names used in R modelling

** Label is true meaning of the variables