PREDICTIVE SCREENING FOR CHRONIC KIDNEY DISEASE

Russell Destremps
Hao Deng
Yue ‘Alex’ Fu
Ravi Bhagwat
Tianyue Wang
Sangyu Zhou

Executive Summary

This report provides insights on developing a prediction model from a dataset comprising of people tested for chronic kidney disease (CKD) and provide a screening tool with **91% accuracy** that can identify individuals with a higher risk of having CKD.

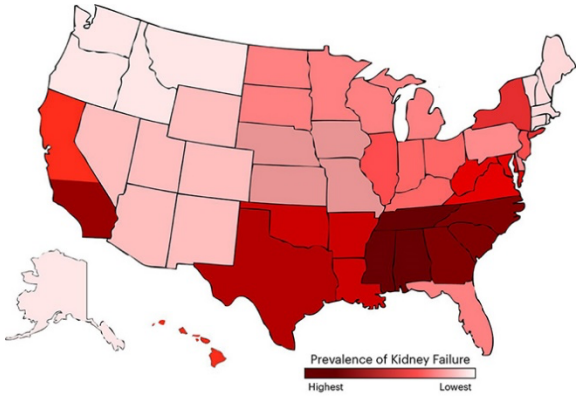
Combined with medical research and statistical analysis, we identified Age, Obesity, Peripheral Vascular Disease (PVD), Hypertension, Diabetes, Cardio-Vascular Disease (CVD), Congestive Heart Failure (CHF), and Anemia are the key variables for identifying the possibility of having CKD. By applying **logistic regression** with a thorough consideration of selecting the **threshold as 0.2**, we defined the weights of each variable providing the scores in the screening tool.

Introduction

"37 million people are estimated to have CKD in United States and 90% with CKD are unaware they have it"

-Centers for Disease Control and Prevention

CKD is a common disease among adults in the United States. According to the latest report[i], 15% of US adults are estimated to have CKD. Unfortunately, most of them do not know they have it. CKD being called as a silent disease slowly affects the proper functioning of kidneys and people might now show symptoms until later stages.



Map showing kidney disease in US states. Source: kidney.org

To improve the situation and help people have an easier and more convenient way to understand their probability of having CKD, we developed a simple screening tool through the identification of critical related variables that can predict whether an individual is identified as being at-risk for CKD based on a 6,000-patient set. This screening tool can then be used to direct high-risk patients to seek more targeted medical screening to confirm the presence of the disease.

The overall objective will then be early informed detection presenting the ability to begin appropriate treatment protocols, potentially resulting in improved quality of life at reduced long-term medical cost.

Data Preprocessing

Through initial analysis, there is one or more variables missing for 31% of the 6,000 patient data set, with complete data for 4136 patient vs missing data for 1864. Since there was a lot of missing data in the observations as well as the variables, we decided to drop observations and variables where there was more than 15% and 5% of missing data respectively. We did it because if we would have tried to impute the missing data, we would have introduced bias in the data. In the process, we deleted 185 observations and 3 variables namely, Income, Poor vision, Unmarried. For the dropped variables, we assumed that people do not want to disclose their income, vision or marital status for obvious reasons but there was no certain way to tell why the data was missing.

After the above process, we were left with 5815 observations and 30 variables where we performed **imputation technique using MICE, library** in R. We used MICE because it imputed on a variable by variable basis and creates multiple imputations as compared to a single imputation which takes care of uncertainty in missing values[ii].

After cleaning and imputing the data, the prediction model needs to be developed. We followed two different approaches to select one at the end based on factors like simplicity, accuracy of the model and several others.

Developing Model

Model 1

The goal of the following model will be to **utilize medical knowledge to develop a model** that identifies variables that either cause or are indicators of the presence of CKD.

Medical research says that CKD is both a major contributing factor and reciprocally a condition that can be developed as a result of various other diseases and medical conditions. These include, chiefly, but not limited to, Hypertension, Dyslipidemia, Diabetes and Cardiovascular Disease[iii]. In either scenario, the presence of one can often lead to the emergence of the other and acceleration of symptoms and progression of both/all diseases. This relationship then creates heightened sense of urgency to recognize and confirm the presence of CKD and enter into appropriate treatment protocol. It is expected then that the presence of CKD in the control observation group will be accompanied by one, if not several of the above conditions.

Moving forward with this medical background, a series of variables will be removed from the fully imputed data set.

- The first set of remaining variables removed were Education, CareSource and Insured. Any one of these variables may intuitively lead to a higher or lower risk of certain medical conditions as they may contribute to higher standards of living, including diet, exercise and availability of quality preventative medical care. Regardless of that, the remaining variables identifying medical stats and disease data, are present and contain the information valuable this analysis.
- Hypertension was retained and Systolic and Diastolic Blood Pressure (SBP/DBP) were dropped, since Hypertension is primarily diagnosed from these measurements[iv].
- Drop Family History of Hypertension and Cardio Vascular Disease, these found to have low correlation in the dataset, .056 and .021 respectively. While Family history of Diabetes was retained, with a .245 correlation to Diabetes.
- Retain Dyslipidemia, drop total cholesterol, HLD and LDL, since Dyslipidemia is primarily diagnosed from these measurements[v].
- Retain obesity, drop weight, height, waist, and BMI. The latter variables are utilized to obtain each other and then used to calculate the Obesity variable.
- Remove Race, research is consistent that higher risk race groups include Pacific Islander, other First Nation Groups. These are not discernable from the dataset, therefore remove[vi].
- The remaining 14 variables following reduction will be brought forward into next phase of analysis, logistic regression. The remaining variables are Age, Obesity, Dyslipidemia, Peripheral Vascular Disease (PVD), Activity, Smoker, Hypertension, Diabetes, Family Diabetes, Stroke, Cardiovascular Disease (CVD), Congestive Heart Failure (CHF) and Anemia.

Logistic Regression

Next, the reduced variables are utilized as predictor variables to run a generalized linear model to explore their interaction with the binary dependent variable, CKD.

	Initial Model *(All variables)*		Final Model *(Most Influential Variables)*		
Variables:	**Coeff.**	**P-value**	**Coeff.**	**P-value**	**Odds Ratio**
(Inercept)	-8.492295	-	**-8.790589**	-	-
Age	0.0880	<2e-16	0.08946	<2e-16	1.09359
Obese	0.2471	0.0489	0.27108	0.02907	1.31138
Dyslipidemia	0.0211	0.9088	-	-	-
PVD	0.6090	0.0004	0.62680	0.00028	1.87161
Activity(Low)	-0.2040	0.1062	-	-	-
Activity(Medium)	-0.1906	0.3560	-	-	-
Acitivity(High)	-0.7211	0.1343	-	-	-
Smoker	-0.0208	0.8587	-	-	-
Hypertension	0.5592	0.0001	0.55919	0.00005	1.74926
Diabetes	0.4679	0.0009	0.45986	0.00060	1.58386
Fam.Diabetes	-0.0898	0.4765	-	-	-
Stroke	-0.0565	0.8296	-	-	-
CVD	0.5769	0.0044	0.56870	0.00023	1.76598
CHF	0.4142	0.0501	0.44125	0.03600	1.55465
Anemia	1.4235	0.0000	1.44217	0.00003	4.22988

Table: Regression summary for model 1

The sign of the predictor coefficients represents whether they have a positive or negative influence on a "1" or positive CKD outcome. The coefficients themselves are in terms of logits, which can be converted to odds ratios by exponentiating each coefficient. To improve the model, variables are successfully removed starting with those with the highest p-values. The resulting model results as follows. This model will next be analyzed via various accuracy measures on the full 6,000 observation training set.

Model 2

As mentioned above on Model 1, Multi-collinearity variables and unrelated variables are removed by studying medical background research after data imputation step. In order to test whether methodology performed in Model 1 is feasible, Model 2 presents a side by side comparison to Model 1 with identical 5815 observation data imputation dataset except removing multi-collinearity & unrelated variables manually on the preprocessing stage. The goal is to find out outcome and compare differences from different angle of interpreting & processing same dataset using **backward variable reduction with multi-collinearity check**.

Linear model with 18 variables remained after backward selection where all P-values are less than 0.05 as shown as Figure 1. However, identifying multicollinearity[vii] within these variables is needed further attention before applying model to logistic regression.

```
Step:  AIC=2100
CKD ~ Age + Female + Weight + BMI + Waist + HDL + LDL + PVD +
    Activity + Hypertension + Fam.Hypertension + Diabetes + CVD +
    Fam.CVD + CHF + Anemia + Racegrpblack + Racegrphispa
```

Fig 1: Regression equation for model 2

```
> car::vif(model3)
        Age      Female      Weight         BMI       Waist         HDL         LDL
   1.529794    2.261870   10.566067    8.280519    6.308887    1.344425    1.124040
        PVD    Activity Hypertension Fam.Hypertension    Diabetes         CVD    Fam.CVD
   1.044934    1.079363    1.134673    2.491422    1.132360    1.182726    2.564139
        CHF      Anemia  Racegrpblack  Racegrphispa
   1.145282    1.055161    1.179005    1.207406
```

Fig 2: VIF values for each variable

Variance Inflation Factor (VIF) function from caret package is applied to check multicollinearity in this case. The result shows that “Weight”, “Waist” and “BMI” have very high VIF value compare to other variables, which and multicollinearity exists variable removal is necessary in such case.

Backward selection is reapplied to the rest of variables and the model 2 result is presented as Figure 3.

In conclusion, 15 of variables are removed from first backward elimination. Through multicollinearity identification, 3 variables, Weight, BMI, and Waist, are removed due to high VIF value. 2 extra variables, race group black and LDL, are removed in the second time of backward elimination.

Model 2 (Final)			
	Coefficients	**P-value**	**Odds Ratio**
(Intercept)	-7.423285	< 2e-16	0.00059718
Age	0.086354	< 2e-16	1.09019208
Female	0.271338	0.029153	1.31171869
HDL	-0.015783	8.48E-05	0.98434067
PVD	0.580637	0.00094	1.78717655
Activity	-0.142494	0.097698	0.86719242
Hypertension	0.593064	2.16E-05	1.80952476
Fam.Hypertension	-0.399489	0.097992	0.67066261
Diabetes	0.46327	0.000809	1.58926231
CVD	0.475268	0.002819	1.60844495
Fam.CVD	0.35366	0.093472	1.42427135
CHF	0.424153	0.048151	1.52829533
Anemia	1.380642	7.46E-05	3.97745412
Racegrphispa	-0.731142	2.30E-06	0.48135883

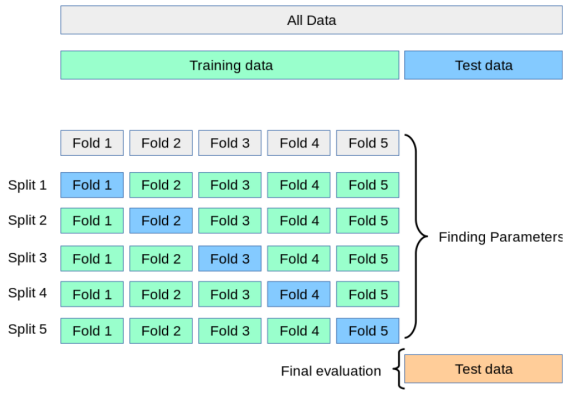
Table: Regression summary for model 2

K-Fold Cross Validation for Model Selection

“Cross-validation (CV) refers to a set of methods for measuring the performance of a given predictive model on new test data sets.”

-Statistical tools for high-throughput data analysis

As a resampling method, CV involves fitting the same statistical method multiple times using different subsets of the data. For a more solid analysis, we picked 10-fold CV in the set of methods to get average values of accuracy, f-measure, and profit among the 10-time training. The primary concept can be shown as below:

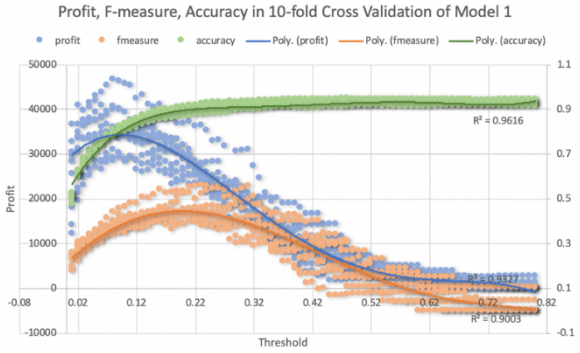


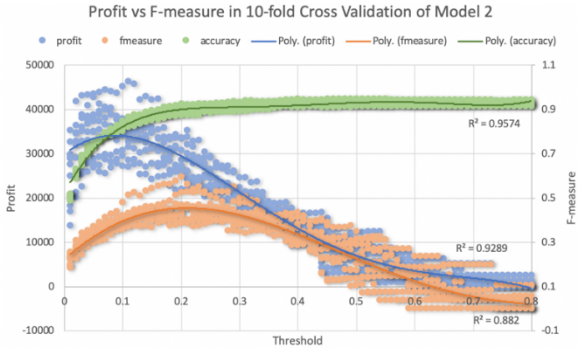
The algorithm of 10-fold CV is as follow:

- Randomly split the dataset to 10 folds.
- Set one of them as test data while other 9 folds as training data.
- Apply logistic regression on the training data getting a set of coefficients of the variables.
- Test the model on the test data and record parameters, which shows accuracy.
- Repeat the process for each subset and record the parameters of evaluating accuracy.

As a better model, besides having a **higher value of primary evaluating criterion**, the value should not vary too much with different training dataset, which means the **variance of the criteria** among the 10 times repeating calculation should be lower. Otherwise, the accuracy would be unpredictable when the model predicts other dataset such as the real test dataset.

For logistic regression, it is crucial to choose a **proper threshold**, which affects the efficiency of the model at a significant level. Therefore, we tried different values of threshold ranging from 0 to 0.8 in the step of 0.01 for each fold of CV. Then we get the diagrams for model 1 and model 2 as below:





As both diagrams showing, profit gets the maximum value at the threshold less than 0.1, f-measure gets the maximum value at threshold is around 0.2, and accuracy doesn't change much after threshold larger than 0.25. To get more precise analysis, we can summarize the maximum value of three parameters at specific threshold in the two models by applying polynomial curve fitting:

Maximum Values at Specific Threshold			
Model	Profit/$	F-measure	Accuracy
1	34711@0.10	0.448@0.2	0.933@0.57
2	34033@0.08	0.445@0.21	0.934@0.56

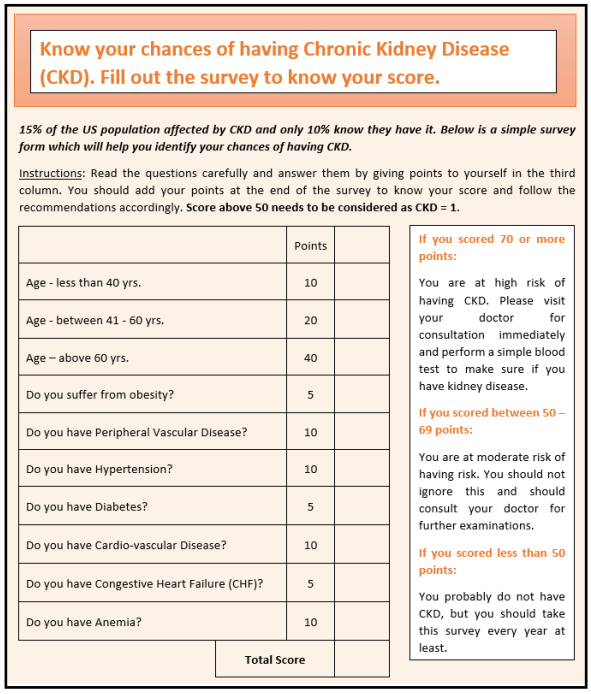
It is worth noting that the profit here is based on 600 observations, which is the size of test data in 10-fold CV.

There is no way of choosing one threshold and making three parameters maximized. To balance the business profit and accuracy, we choose **f-measure as the primary criterion**. Therefore, threshold 0.2 and 0.21, where f-measure is the maximum value of the model 1 and 2, respectively, would be chosen as the **final threshold**. The final performances of the two models are shown below:

Final Performance			
Model	Profit/$	F-measure	Accuracy
1w/0.2	29130	0.448	88.94%
2w/0.21	28250	0.451	89.49%
Standard Deviation			
Model	Profit	F-measure	Accuracy
1w/0.2	3853	0.049	0.014
2w/0.21	4447	0.059	0.017

As we can see, the performance of the two model is slightly different. Model 1 has higher profit and lower standard deviation for all parameters, which means model 1 will act more consistently to different dataset. Also, for the simplicity of making a tool, we choose model 1 as the final model.

Simple Survey Tool



Know your chances of having Chronic Kidney Disease (CKD). Fill out the survey to know your score.

15% of the US population affected by CKD and only 10% know they have it. Below is a simple survey form which will help you identify your chances of having CKD.

Instructions: Read the questions carefully and answer them by giving points to yourself in the third column. You should add your points at the end of the survey to know your score and follow the recommendations accordingly. **Score above 50 needs to be considered as CKD = 1.**

	Points	
Age - less than 40 yrs.	10	
Age - between 41 - 60 yrs.	20	
Age – above 60 yrs.	40	
Do you suffer from obesity?	5	
Do you have Peripheral Vascular Disease?	10	
Do you have Hypertension?	10	
Do you have Diabetes?	5	
Do you have Cardio-vascular Disease?	10	
Do you have Congestive Heart Failure (CHF)?	5	
Do you have Anemia?	10	
Total Score		

If you scored 70 or more points:

You are at high risk of having CKD. Please visit your doctor for consultation immediately and perform a simple blood test to make sure if you have kidney disease.

If you scored between 50 – 69 points:

You are at moderate risk of having risk. You should not ignore this and should consult your doctor for further examinations.

If you scored less than 50 points:

You probably do not have CKD, but you should take this survey every year at least.

As our one of our objectives, we created a simple survey tool which would make people aware about their chances of having Chronic Kidney Disease. Since we chose Model 1 for our analysis, we developed a tool with points assigned to each question according to their coefficients from the regression equation and then scaled to be simple and easy to understand by everyone.

One of the challenge was to deal with one continuous variable, age. In order to assign the points to age, we calculated the percentage of people affected by CKD in different age ranges. We found out that one 3% of people below the age of 40 are affected by CKD. Only 9% of people between the age of 41 – 60 are affected by CKD and more than 25% of people above the age of 60 are affected by CKD. Several other age ranges were also considered, however, the above was found more reliable for the survey.

Conclusion

The possibility of having Chronic Kidney Disease (CKD) should be checked regularly in adults. People above the age of 60 should be at high alert and who are affected by diseases like Hypertension, Cardio-Vascular Disease, Diabetes and several others which we figured out in our analysis.

We believe that incorporating medical research and data relative to CKD, helped us strengthen the accuracy of our model and also increased the chances of our analysis and tool being accepted by the medical professionals.

Limitations

- The dataset provided to us is not being completely representative of standard U.S. population.
- The size of our training dataset was not large enough, which might have lowered the accuracy of our model.

Citations

[i] Centers for Disease Control and Prevention. *Chronic Kidney Disease in the United States, 2019*. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2019.

[ii] *https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/*

[iii] Lipman, Mann & Schiffrin. Chronic Kidney Disease: Effects om the Cardiovascular System.

[iv] How High Blood Pressure is Diagnosed. Retrieved from: *https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/how-high-blood-pressure-is-diagnosed*

[v] Ishwarlal Jialal, MD, PHD, MRCPATH, DABCC, A Practical Approach to the Laboratory Diagnosis of Dyslipidemia, *American Journal of Clinical Pathology*, Volume 106, Issue 1, 1 July 1996, Pages 128–138, https://doi.org/10.1093/ajcp/106.1.128

[vi] Pfeifer, Phillip E & Reynolds, Richard S. (March 21, 2013). *Screening for Chronic Kidney Disease.* Darden Business Publishing, University of Virginia.

[vii] Identifying Multicollinearity in Multiple Regression, from: http://www.researchconsultation.com/multicollinearity-regression-spss-collinearity-diagnostics-vif.asp