C.K.D Project

STAT-630-674

Group 3

Anqi Wang

Apurva Tawde

Chenyang Liu

Yibin Song

Tu Trinh Nguyen

Executive Summary

Chronic kidney disease includes conditions that damage the kidneys and decrease their ability to keep a person healthy. The aim of this research was to develop an easy screen questionnaire with acceptable accuracy that is able to detect individuals at high risk of CKD before expensive confirmatory laboratory testing.

The Report consists of data analysis and exploration, and methods of approach and what are the important factors, in order to design the easy screening tool. This will help early detection and treatment which often keeps chronic kidney disease from getting worse. Method of analysis includes Principal Components Analysis, Factor Analysis, and Cluster Analysis. Results of data analyses show that these important variables are associated with a higher chance of getting CKD: **Age, LDL, Diabetes, Hypertension, Anemia, SBP, Cardiovascular Disease**

This report concludes that these variable are the best for being the predictor of CKD. Through this report, there is going to be an explanation of how the screening tool works and prove its accuracy.

Who is highly at risk?

- **Age**: People 60 years and older are at a higher risk of developing CKD.
- **LDL :** people with low- density lipoproteins level leads the buildup of cholesterol.
- **Diabetes**: It is a disease that occurs when the blood glucose is too high. It is an important cause of CKD.
- **Hypertension:** also called high blood pressure, is the second-highest cause of CKD.
- **Anemia:** lack of enough healthy red blood cells or hemoglobin in the blood causes Anemia. Patients with CKD are also diagnosed with Anemia over a period of time.
- **SBP:** Increase in SBP is significantly associated with CKD. SBP may be the most clinically useful predictor of CKD.
- **Cardiovascular disease**: In addition to hypertension, other diseases of the heart and blood vessels may increase your risk for kidney disease.

Objective

The main objective is to select variables out of **33 variables** to predict the value of CKD as accurate as possible, and create an **easy screen tool** based on the analyses. This easy screen tool is necessary because when kidney disease progresses, it may eventually lead to kidney failure, which requires dialysis or a kidney transplant to maintain life.

Therefore, **an easy screening tool** may help to address the burden of Chronic Kidney Disease **(CKD)**. After **cleaning the data** and **replace the missing values**, the data was tested by using different approaches in order to identify the key variables that would highly be associated to diagnose CKD. **A simple screening tool** for patients can easily access before they have to go through the expensive testing procedure.

Data

The Raw Data

There are **34 variables** in the data set.

- ➔ From column A to column I are demographics in nature
- ➔ From column J to column U were the results collected during the physical exam
- ➔ From column V to column AH is based, in part, on self-reported health histories

These variables are easily obtained by a family physician during routine checkups except for cholesterol and hemoglobin count require a blood test.

This data **is not** a **random sample of U.S. adults**. Since the data is not missing at random, for a better prediction, it is necessary to replace the missing values.

Replacing Missing Value and Variable Deduction

When considering the entire dataset of patient records, out of all variables, **missing records** present in 23 independent variables.

- ❖ For **RaceGroup**, with external research, African American and Hispanic have a higher chance to get CKD. Therefore for these two ethnics, the value will be 1 and else will be 0.
- ❖ **Missing Binary Variables:** Hypertension, Diabetes, Anemia, CVD, and Obese, if missing, are given values based on the results of **logistic regression**. If still missing, it is assigned the mean of the variable.
- ❖ **Numeric Variables:** SBP, DBP, LDL, Weight, and Height were replaced with Mean after running **Linear Regression** but still missing.

- **Grouping:** To reduce the number of variables from the beginning, based on the case study and other medical reviews, CVD, Stroke, CHF and PVD are **grouped** together as a Cardiovascular disease group. So is BMI Group (with Weight, Height, Obese and BMI). If either of them has a value of 1, the value of the new group gets the value of 1, otherwise 0.

Analysis

The Variable considered after the data cleaning step:

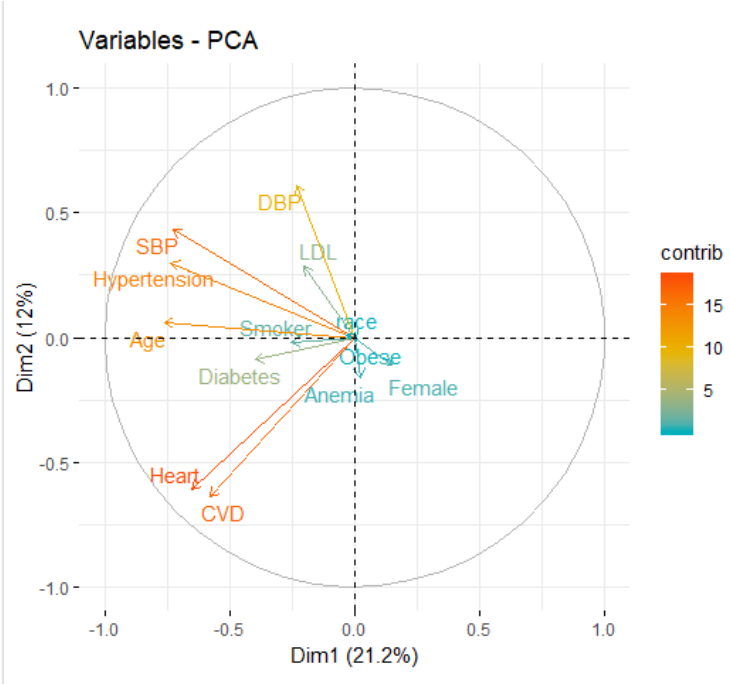
- Age, Female, Race
- Obese
- Blood pressure (including SBP and DBP)
- Cardiovascular disease (based on the value of PVD, Stroke, CVD, and CHF)
- Smoker
- Hypertension
- Diabetes
- Anemia
- LDL

For Dimension Selection, the following **3 methods** were selected:

1. PCA

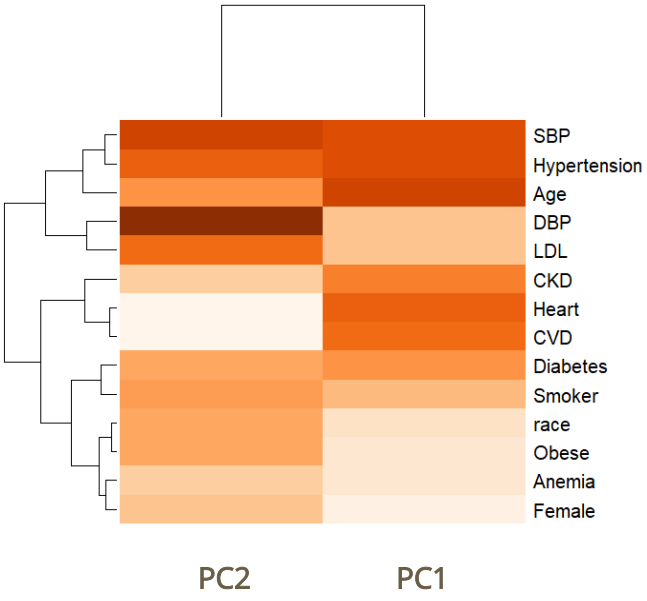
- According to the PCA, the high contribution are Age, Hypertension, Cardiovascular Disease, SBP, and DBP. Refer to **Appendix 1** for the Eigenvalue, the first five PCA were chosen, because of their **eigenvalues** larger than **1**.

- As the most important PCA, **PCA1** represents Age, Hypertension, SBP. **PCA2** represents DBP and



Cardiovascular Disease. **PCA3** represents Female. **PCA4** represents Race. **PAC5** represents Smoker.

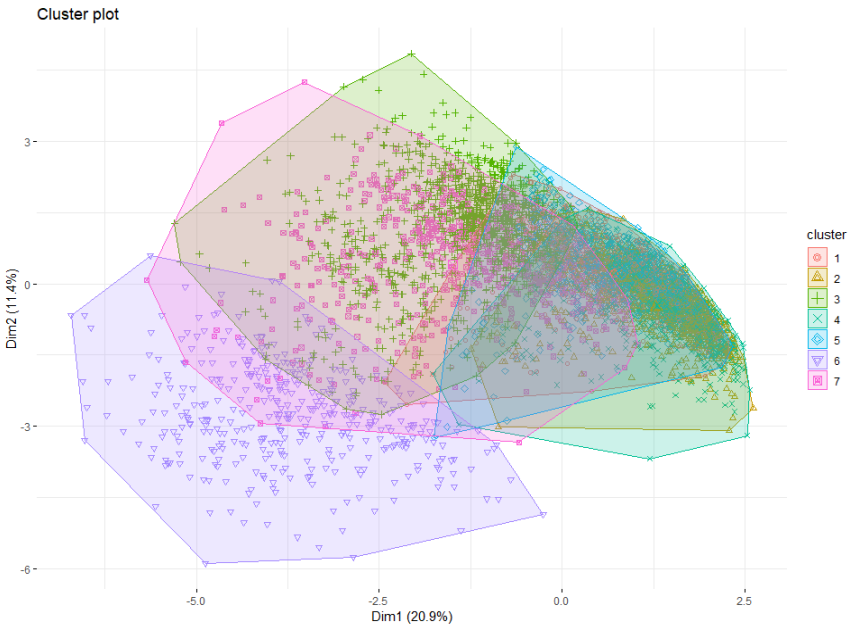
2. Factor Analysis



- From Factor analysis and the Heat Map, **PC1** associates with Age, Hypertension, Cardiovascular Disease, CVD, SBP. **PC2** associates with DBP, LDL, SBP, and Hypertension.

3. Clustering using K-Means

Clusters	Age
1	-0.6139557
2	0.4844676
3	-0.1378993
4	-0.4335648
5	0.8888559
6	1.2855748
7	0.4705895



- When the clusters were compared with the CKD column, 2 of high clusters had CKD value
- As from the data it was evident that cluster **6** & **2** were **high** in CKD, the other variables in the cluster that was high are for **Cluster 6**: **Age, Diabetes, Hypertension, SBP & Cardiovascular Disease.**
- Also as per our observation in **Cluster 2**: **Anemia & Female(Gender)** were high when CKD was high

Logistic Regression

- Based on careful observation and results of **PCA**, **Factor analysis** and **Clustering,** we chose 13 variables. After using backward stepwise logistic regression, we finally determined 7 variables in our model.
- Below is the Model Generated which is used for CKD Prediction

.5 + 0.09 x Age + 0.003 x LDL + 0.53 x Diabetes + 0.74 x

Screening tool

Additional research more on what symptom and additional information would help to better identify and make the screening tool simple. Refer to **Appendix 2** for the screening tool.

- Age
- LDL
- Diabetes
- Hypertension
- Anemia
- BMI(>30)
- Cardiovascular Disease

Conclusion

As mentioned above, with three variable deduction methods, the number of independent variables are narrowed down to 13. Based on the comparison of actual CKD and predicted CKD based on the logistic regression, the accuracy with this model is acceptable to create an easy screen tool. The total accuracy is **77%**, based on the original **6000 records** (including **407** true positive results, **4214** true negative results, **1322** false positive results, and **57** false negative results).

References

- Centers for Disease Control and Prevention. National Chronic Kidney Disease Fact Sheet, 2017. National Chronic Kidney Disease Fact Sheet, 2017 (PDF, 1.32 MB) . Accessed June 7, 2017.
- Gheewala, Pankti A et al. “Public knowledge of chronic kidney disease evaluated using a validated questionnaire: a cross-sectional study” BMC public health vol. 18,1 371. 20 Mar. 2018, doi:10.1186/s12889-018-5301-
- Harward, Donna H et al. “Evaluation of the Scored Questionnaire to Identify Individuals with Chronic Kidney Disease in a Community-based Screening Program in Rural North Carolina” Journal of community medicine & health education vol. 4,Suppl 2 (2014): 007.

- Race, ethnicity, and kidney disease. NIDDK website. www.niddk.nih.gov. Published June 13, 2017. Accessed June 13, 2017.

Appendix

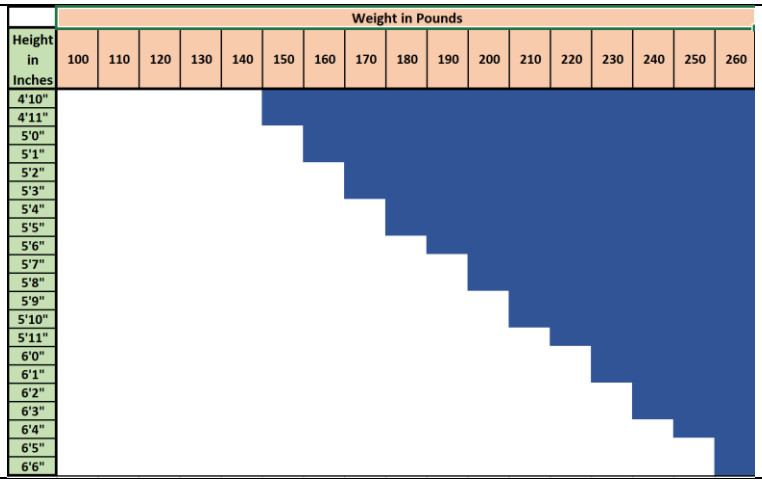
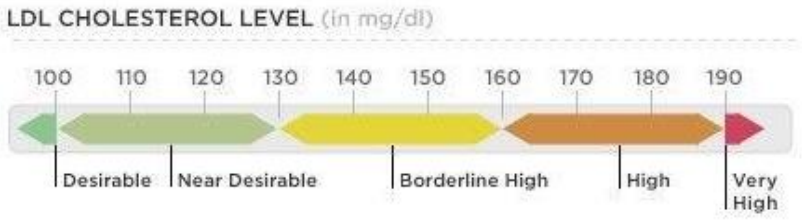
Appendix 1

```
> eigenvalues
 [1] 2.7583042 1.5651087 1.1995325 1.0810728 1.0380594 0.9690497 0.9302872 0.9131299 0.8540175 0.6888126 0.4413296 0.3330111 0.2282848

> eigenvectors
              [,1]         [,2]        [,3]        [,4]        [,5]
 [1,]  0.458733980 -0.049784519 -0.09375659 -0.08051061  0.18340198
 [2,] -0.090125018  0.089513121 -0.68154590 -0.03703530 -0.08124746
 [3,]  0.140989209 -0.489595863  0.16086156  0.17081685 -0.36593463 -
 [4,]  0.123538202 -0.228339650  0.14865892 -0.16017118  0.01010312
 [5,]  0.152135328  0.013785320  0.40922794 -0.21612548  0.57905693 -
 [6,]  0.239766031  0.072023169 -0.14169413 -0.12725389  0.40215453
 [7,]  0.443611511 -0.240368828 -0.22316519  0.01743163 -0.01320381 -
 [8,]  0.349529171  0.509572683  0.13856062  0.10123937 -0.28306537 -
 [9,] -0.014173190  0.127815455 -0.42697130 -0.05219499  0.27443307 -
[10,]  0.438200819 -0.347078784 -0.17765786  0.06512278 -0.05479825 -
[11,]  0.021209928 -0.008965994  0.01303780 -0.66460536 -0.23601657
[12,] -0.003138724  0.009109346 -0.02278646 -0.63904716 -0.25159733 -
[13,]  0.392461080  0.487732601  0.08732969  0.08251871 -0.22604293 -
```

Appendix 2

No.	Questions	Yes/No	Points
1	I am in the age of 45 to 59	Yes	2
2	I am in the age of 60 to 70	Yes	3
3	I am in the age of 71 and older	Yes	4
4	I have/had Anemia	Yes	1
5	I have/had High blood pressure (Hypertension)	Yes	1
6	I am diabetic	Yes	1
7	I have a history of Cardiovascular disease (including heart attack, stroke, congestive heart failure, coronary artery disease, and peripheral vascular disease)	Yes	1
8	From the below picture in which zone you identify yourself		White=0 Blue=1

			
9	Have you ever tested for Cholesterol? 		2= Have Cholesterol (Orange/Red) 1=Borderline High (Yellow) 0=Don't have
Total Points			

- If the total points are from **0-4** points, then CKD **is not positive** but should consider retaking the screening tool again in the future

- If the total points are above **4** points, the patient has a **higher chance** to get CKD so they should go take the blood test or CKD test for further identification.