SCREENING FOR CHRONIC KIDNEY DISEASE

STAT 630 *- Report date: Feb 25, 2020*

Group 6
Neel Kumtakar
Daniel Le
Yusi Luo
Kim-Cuong Nguyen
Michael Sadi
Jing Wang

INDEX

EXECUTIVE SUMMARY

The team was tasked with creating an easy-to-use screening tool to identify people who run a high risk of CKD so that they can get treatment early, given the clear benefits of early detection and treatment.

As part of the analysis, we employed a clinical dataset gathered by the National Center of Health Statistics of 6,000 adults aging 20 years or older from 1999-2002. The team ran logistic regression on this dataset a few times to get the final predictive model on probabilities of CKD presence at patients. We first applied backward elimination approach to get the first model, from which we analyzed and defined significant variables that should be included in the screening tools. By narrowing down the number of variables to 9, we ran a reduced model and compared its performance to the first model – the backward elimination one.

We concluded that the reduced model got similar performance with the backward model and decided to go ahead with it to create the screening tool. We used a cutoff of 0.22 for prediction, yielding $71,900 with the accuracy level of 90%. We noticed that the classification by this model is highly skewed towards negative results. The model precision is low – 36%, which means that we were correct in 36 out of 100 cases when we predicted CKD presence (positive result). However, in terms of our failure to predict CKD presence (false negatives), the percentage is 2.5% only. Thus, from a medical perspective, we still could help most of the patients with CKD to detect their diseases early.

The screening tool was created based on the 9 variables utilized in the predictive model. The tool is designed as a questionnaire with 9 questions relevant to the 9 variables. Each question is assigned with different points. The screening tool could yield a profit of $49,600 at 90% accuracy.

Our analysis still contains several limitations that we need to work on, including the drawback of multiple imputation for missing data, drawback of backward elimination model, the incomplete dataset which does not have information on family history of kidney disease – an important CKD risk factor, and potentially incorrect input data provided by the patients included in the dataset.

INTRODUCTION

Chronic Kidney Disease (CKD), according to the Mayo Clinic, is a disease that describes the decay of the kidneys. This kidney disease is gaining growing public awareness for its severe impacts. People with this disease are normally nauseous, vomit a lot, have trouble sleeping, and other major problems. Due to the importance of kidney function in maintaining homeostasis, CKD can affect almost every body system. Each year, CKD has killed more people than other diseases because of no detection in the early stage. Early detection and treatment are essential to slowing the disease progression and improving outcomes, ensuring a person's health status and life quality.

According to some previously conducted research, some risk factors that may increase the probability of kidney problems.

- Age: People over 60 have higher risk of CKD.

- Race: white people have higher risk.

- Diabetes, Hypertension, CVD, family history of kidney disease: people having the mentioned diseases, or a family member used to suffer from kidney problems run higher CKD risk.

OBJECTIVE

The purpose of this case study is to develop a simple, easy-to-use screening tool to identify people who run a high risk of having CKD, so that treatment can be applied early enough, given the clear benefits of early detection and treatment. The team employed a clinical dataset gathered by the National Center of Health Statistics of 6,000 adults aging 20 years or older from 1999-2002 through surveys, examinations, and blood samples, based on which we analyzed and defined key factors that could indicate the presence of CKD. These factors were utilized to create a simple, easy-to-use screening tool. The dataset includes 33 variables, ranging from demographic information like age, gender and race to pre-existed health concerns like cholesterol, stroke, PVD, etc. The key method in this analysis is logistic regression, by which we could classify if a patient has CKD or not.

INITIAL ANALYSIS

A quick analysis was performed on the dataset to see the common patterns within each of the two groups - people having CKD and people not having CKD.

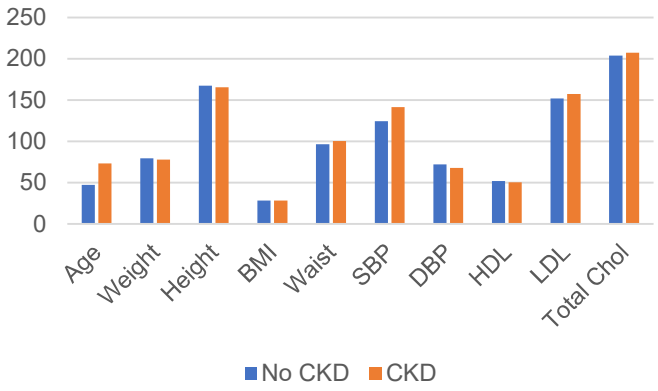

Figure 1: CKD dataset – Continuous variables

The first chart gives an overview on the mean values of continuous variables concerning two mentioned groups. The variables of concern are comprised of age, physical traits (including weight, height, waist), blood pressure level, and cholesterol level. Some initial findings are:

- Older people run higher risk of CKD. The average age of the group of CKD patients is 73 while the corresponding value of the group without CKD is 47.

- High blood pressure can be one of potential indicators of CKD presence. CKD patients have

an average blood pressure of 141 mm Hg while people without CKD have an average of 124 mm Hg. The normal range of SBP is said to be lower than 120 mm Hg. Thus, it can be said that people with CKD tend to have high blood pressure.

- There is no clear difference between CKD patients and no-CKD people in terms of physical traits and cholesterol levels.

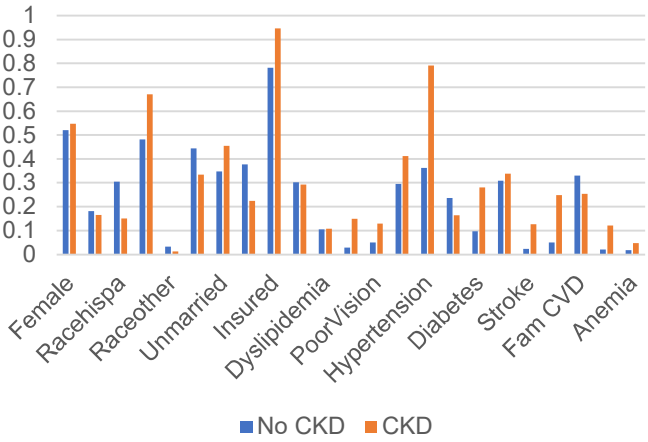

Figure 2: CKD dataset – Binary variables

The second graph shows comparison between people with and without CKD. The columns represent the percentage of people with/without CKD with such features. Interestingly,

- It seems that white people run a higher risk of CKD while a Hispanic has a low risk. The percentage of Hispanic without CKD doubles the figure for Hispanic CKD patients.

- People with more education and higher income less suffer from CKD. This can be explained by the fact that these people are better equipped with health knowledge and know to maintain better health conditions.

- Hypertension seems to lead to CKD when 80% of CKD cases have hypertension. This well corresponds to the finding above on high SBP average among CKD patients.

- PVD, PoorVision, Diabetes, Stroke, CVD, CHF can also be inferred as indicators of CKD presence. There is a sharp contrast between the percentage of patients with and without CKD with these symptoms.

- Surprisingly, though such diseases can be indicators of CKD presence, having a family history of those disease are not.

MODEL FITTING

Missing Data & Imputation

The dataset of concern has 33 variables and 6,000 observations in total, of which 1,864 observations, accounting for 31%, have missing values. The variable “Income” has the highest number of missing values (13.2%). The missing value in income is highly likely to be missing not at random (MNAR), which means there is a process behind the generation of the data that influence the missing values; for example, people earning higher income prefer not to reveal their income status. Therefore, if we choose to handle missing value with the method of listwise deletion, there is a high likelihood that the result would be biased.

There are a number of common ways to handle missing values; for example, listwise or pairwise deletion, dropping variables, imputation with mean, median, and mode, and multiple imputation. In this case, we decided to use multiple imputations to replace missing data in the CKD dataset for the following reasons,

- Missing data are distributed widely throughout the dataset and do not concentrate on some specific variables only.

- Missing value in some variables seem to be MNAR.

- The larger the dataset is, the better we can train our model.

The imputation process was based on linear regression for predicting continuous variables and logistic regression for categorical variables. In order to impute the dataset, we utilized package MICE.

Model Development

Once we got our complete imputed dataset, we randomly split the dataset into training data and validation data by 75%-25%. We chose 75% for training data because we believe that the dataset with 6,000 observations is big enough and 75% of the dataset is good enough to train our model.

Our purpose of model development here is to figure out a model to predict probabilities of CKD presence. Based on the model, we will utilize the model's independent variables to create an easy-to-use screening tool. Concerning the screening tool, we set two priorities as below:

- ***Simplicity:*** The screening tool should be simple enough so that all the patients could understand and employ. We aim to reduce the number of variables in the tool as much as possible and give more priority to binary variables to ensure the ease of using the tool. We keep in mind that this should be a tool that our grandparents could well understand and use as well.

- ***Accuracy*****:** is a must when developing the tool. It obviously makes no sense to use a tool which wrongly predicts CKD presence.

Considering the above priorities, we did not run logistic regression on the full 33 variables but decided to apply backward elimination approach to reduce the number of variables in the model.

When it comes to selection process for multiple regression, there are two approaches of forward selection and backward elimination. With forward selection, we start with a null model and add a new feature at each step. This approach brings a drawback when the addition of a new variable can possibly eliminate the significance of the already included variables. Therefore, we believe backward elimination is a better approach where we start with a full model and step by step omit insignificant variables.

Backward Elimination Model

Figure 3 shows the result of the backward elimination model. Significance level is 0.05.

	Estimates	p-Value	Significant
Intercept	-7.19608	< 2e-16	
Age	0.092035	< 2e-16	Yes
Female	0.414177	0.004247	Yes
Racehispa	-0.54453	0.017292	Yes
Raceother	-0.44584	0.427289	
Racewhite	0.180947	0.323975	
Weight	0.020208	0.015931	Yes
Waist	-0.02311	0.027792	Yes
HDL	-0.0178	0.000112	Yes
PVD	0.652613	0.001218	Yes
Activity	-0.15152	0.119707	
PoorVision	0.355674	0.044827	Yes
Hypertension	0.587798	0.000219	Yes
Fam.Hypertension	-0.45651	0.079436	
Diabetes	0.458562	0.003882	Yes
CVD	0.43438	0.012819	Yes
Fam.CVD	0.490381	0.028958	Yes
CHF	0.514173	0.026887	Yes
Anemia	0.995279	0.015573	Yes

Figure 3: Backward Elimination Model Summary

At a quick glance, significant variables in this model are well aligned with our initial analysis.

Reduced Model

As mentioned before, the main objective is to develop an easy-to-use screening tool for CKD presence. We also set two priorities for our tool. Accordingly, the backward elimination model is still not a good choice. Therefore, we took one more step to create another model with less variables. We selected variables for the reduced model by picking the significant variables in the backward model only and omit the variables that we consider will negatively impact the simplicity of the screening tool. We left out the following variables despite their significance:

- Weight, Waist, HDL: The information may not be available at the time when patients use the screening tool, especially HDL, which is the patient's good cholesterol and thus needs further testing to get results.

- Poor Vision: this is self-reported poor vision. There is no clear definition on how poor the vision should be so that patient can properly report this.

So, we were left with 9 variables only, including Age, Female, Race, PVD, Hypertension, Diabetes, CVD, CHF, Anemia. The model summary is shown in Figure 4. Significance level is 0.05.

	Estimates	p-Value	Significant
Intercept	-8.39449	< 2e-16	
Age	0.08524	< 2e-16	Yes
Female	0.15888	0.216967	
Racehispa	-0.61735	0.003971	Yes
Raceother	-0.60266	0.277479	
Racewhite	0.12679	0.475121	
PVD	0.65235	0.00099	Yes
Hypertension	0.57072	0.000249	Yes
Diabetes	0.57757	0.000135	Yes
CVD	0.54826	0.00132	Yes
CHF	0.57249	0.012508	Yes
Anemia	1.0099	0.012477	Yes

Figure 4: Reduced Model Summary

Comparison

We made a comparison between the backward selection model and the reduced model to check how much the model performance is impacted by the reduction of variables. Accuracy was conducted at cutoff level of 0.5.

	Backward Model	Reduced Model
Null deviance	2523.2	2523.2
Residual deviance	1678	1717.4
AIC	1716	1739.4
Accuracy	93%	93.3%
TP	15	14
TN	1381	1385
FP	18	14
FN	86	87

Figure 5: Comparison Between Two Models

As we can see from the table, the backward model is more significant because the reduced model has higher AIC (generally an estimator of prediction error) and the gap between null deviance and residual deviance of the reduced model is wider (the gap shows how well the model performs against the null model – model with intercept only). However, considering approximate accuracy among the two models, we decided to go with the reduced model and made more analysis on it.

Evaluation of the Reduced Model

From the two plots below of the reduced model, we could easily realize a tradeoff between Accuracy and Profit. Profit decreases while Accuracy increases when cutoff value approaches towards 1.

We decided to choose the cutoff value of 0.22, which was supposed to yield an accuracy level of 90%. Our rationale behind this choice is: One of

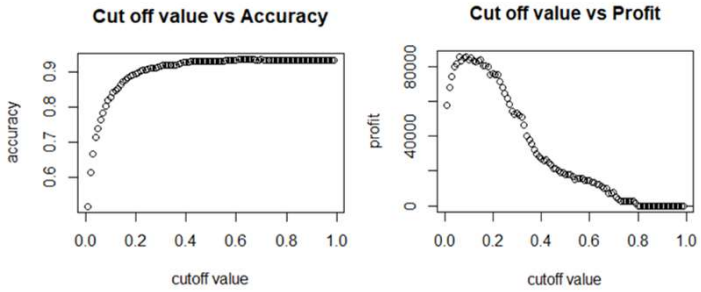

Figure 6: Cutoff value vs Accuracy & Profit

the business goals is to maximize profits; we could get $1,300 for a True Positive and are penalized $100 for a False Positive. However, business cannot go sustainable without a good reputation. Therefore, we did not want to make a big tradeoff between profit and accuracy. We believed 90% is a reasonable accuracy level in this case.

With the cutoff value of 0.22, we got the confusion matrix below when applying this model to the validation data of 1,500 observations. The yielded profit in this case is $71,900.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Value	Positive (1)	64	113
	Negative (0)	37	1286

Figure 7: Confusion Matrix

From this confusion matrix, it's easily seen that classification is highly skewed – the model says no most of the time. The model precision is low, just 36%. This means that out of 100 cases that the model says yes to CKD, only 36 cases are correct. This is surely a drawback of the model. However, considering our purpose of helping patients to early detect and treat CKD, this is not quite a problem anymore. What truly matters from a medical perspective is if we fail to predict in case of CKD presence. Although we wrongly predicted 7.5% of cases as positive though they don't have CKD, the number of cases when we couldn't predict CKD presence was 37 out of 1,500, accounting for 2.5%.

SCREENING TOOL

We utilized the 9 defined variables to create out screening tool (Appendix 2). The screening tool is a short questionnaire for patients to fill out. Each answer will be assigned different scores based on the importance. The score range is from 0 to 13; we set the cutoff points to be 8. If a person's total score is 8 or higher, he/she would need to undergo more testing. Applying the screening tool to validation data, we got $49,600 at an accuracy level of 90%.

LIMITATIONS

- Bias from data imputation: We used multiple imputations to replace missing value. This process potentially lead us to bias interpretation.

- Incomplete dataset: Research has shown that family history of kidney disease is one of the risks of CKD presence. However, this variable is not available in our dataset.

- Backward selection model drawback: The models selected by this procedure may include variables that are not necessary. Thus, we could try to use more different procedures to select variables to achieve more accurate predictions.

- Incorrect input by patients: Patients may give incorrect data. They may not want to disclose some sensitive information, or they don't really know that they have some diseases. That's why the model can only identify the common patterns among patients within each group of having or not having CKD and give predictions but cannot give 100% correct predictions.

APPENDIX 1 – Variable Definitions

Variables	Definition
ID	Identification
Age	Age in Years
Female	Gender, 1 if female, 0 if Male
Racegrp	Ethnicity, white, black, hispanic, other
Educ	Education Level, 1 if more than HS
Unmarried	1 if unmarried
Income	1 if household income is above the median
CareSource	Self reported source of medical care (Dr, HMO, Clinic, nonplace, other
Insured	1 if covered by health insurance
Weight	Weight in KG
Height	Height in CM
BMI	Body Mass Index in kg/m^2
Obese	1 if BMI is above 30
Waist	Waist Circumference
SBP	Systolic Blood Pressure
DBP	Diastolic Blood Pressure
HDL	good cholesterol
LDL	bad cholesterol
Total Chol	HDL + LDL
Dyslipidemia	Too high LDL or too low HDL
PVD	Peripheral vascular disease reflected by reduced SBP at leg relative to arm
Activity	1, 2, 3, 4. The greater the number, the greater the activity
Poor Vision	Self reported self vision
Smoker	whether the patient has smoked at least 100 cigarettes or not
Hypertension	Presence of at least 1 of 4 blood pressure indicators
Fam.Hypertension	family history of hypertension
Diabetes	self reported physician diagnosis or lab test result
Fam.Diabetes	family history of diabetes
Stroke	Self reported, whether the patient had a stroke or not
CVD	Self reported, whether the patient had cardiovascular disease
Fam.CVD	family history of cardiovascular disease
CHF	Self reported, whether the patient had Congestive heart failure
Anemia	Whether the patient had Anemia or not
CKD	Whether the patient had Chronic Kidney Disease

APPENDIX 2 – Screening Tool

1) General Information

Doctor's use only

What is your age?	Below 50 (0)	50-59 (3)	60-69 (4)	Above 70 (5)
	☐	☐	☐	☐
What is your gender?	Male (0)	Others (0)	Female (1)	
	☐	☐	☐	
What is your race?	Black (0)	Hispanic (0)	Others (0)	White (1)
	☐	☐	☐	☐

2) General Health Information

Do you have Peripheral Vascular Disease (PVD)?	No (0)	Yes (1)
	☐	☐
Do you have hypertension?	No (0)	Yes (1)
	☐	☐
Do you have diabetes?	No (0)	Yes (1)
	☐	☐
Do you have Cardiovascular disease (CVD)?	No (0)	Yes (1)
	☐	☐
Have you ever had Congestive Heart Failure (CHF)?	No (0)	Yes (1)
	☐	☐
Are you anemic?	No (0)	Yes (1)
	☐	☐

Total: …/13

Instructions: If total point is 8 or higher, the patient need to undergo more testing.