Screening Tool for Chronic Kidney Disease

Presented to

Dr. Matthew Schneider

February 2019

Group 3

Vinay Gandhi

Ryan Ferguson

Teng Shuai

Sisi Fu

Qintian Qi

Executive Report

Introduction

Chronic Kidney Disease (CKD) refers to the slow deterioration of kidney function. The problem with this disease is that the symptoms will only be obvious on a more advanced stage disorder. By using statistical approach to predict CKD on earlier stages, doctors would be able to diagnose the problem earlier and thus resolve it before it goes to a more advanced stage.

There are two main causes of chronic kidney disease, diabetes and high blood pressure, which are responsible for up to two-thirds of the cases. Cardiovascular disease also increases the risk for kidney disease. Furthermore, age is considered as an important factor to test if a person has CKD, because along with the increase in age, there is an increase in the risk for kidney disease. Based on the case study, we also learn that race is a factor that we need to take into consideration, because people belong to certain ethnic groups are at higher risk for developing chronic kidney disease.

After checking the 33 variables, we found out that CVD, Stroke and CHF are diseases of the heart and blood vessels which may increase the risk for kidney disease. Then we did correlation between these three variables and found that they are highly correlated to each other. In order to make each variable in our model as independent as possible, we only choose one of these three.

Model

To predict the decision of a person having CKD we build different models using logistic regression. With the help of this we will get probabilities and hence say if he/she has CKD (CKD = 1) or doesn't have CKD (CKD = 0). Model will be based on the factors that most affect a person having CKD. As it was mentioned earlier the two major causes of CKD is diabetes and hypertension. These two variables will be of high significance in any model and hence will be selected. In terms people who have risk of getting CKD factors like diabetes, hypertension, Cardiovascular disease, family history of kidney disease, age and race are determined, however we consider only those who data is available.

For us to find a model more efficient and accurate we add, drop some variables and our final model had these variables namely Age, Hypertension, Diabetes, Cardio Vascular Diseases and Race group. In the dataset we found out that people with white race group are more likely to have CKD hence create a dummy variable for people with white race and use only that for our model. Age is most significant variable in our model in terms of data. All the variables in the model have very low p-value, low p-value suggest that the variable is significant in a model hence it is very important that all the variables we have selected have a low p-value. We built different models and different variables and because of high p-value weren't selected, which will be discussed below

To check the accuracy of our model we split the dataset of known CKD into train (2/3 of the data) and test (1/3 of the data). After that we calculated the accuracy of our model based on true positives, true negatives, false positive and false negatives. True positives value indicates number of people who has CKD and we predicted that they have CKD. Another important factor is false positive which says the amount of people we predicted have CKD but don't have it. Accuracy is measured with ration of true positive and sum of true positive and false positive.

Other Models

Model one: Age+PVD+PoorVision+Diabetes+CVD+Hypertension+Racegrpwhite+Anemia

Model two (Our Choice): Age+Racegrpwhite+Diabetes+CVD+Hypertension

A good model is the one that we can test more CKD in a high probability compared with other models. It is very important to find a proper probability to test the model. 25% to 40% is a good range of probability since we can see obvious difference in the number of TP between each model Also, we can maintain keep a large sample at the same time. If we decrease the test probability significantly, the number of CKD in each model will be very close based on the 4136 complete data. For example, if we choose 7% as CKD, nearly every model will test similar number of CKD, which is almost 230. Also, when we choose 50% as the probability of CKD, all the model will get similar result too.

When we choose 25% as CKD, Model one can test 119 CKD while model two test 112 CKD. That means model 1 have a more accurate result when the we choose the probability of CKD is above 25%. When we choose 40% as CKD, the difference will be much more obvious. The number of CKD is 51 in model one while the number in model 2 is 42. However, if we split 6000 data into 2 parts, one is 4000 and another one is 2000. When we test it in regression, the accuracy of the first model decreased while the second model's accuracy is almost the same. So, the second model is more stable than the first one. Besides, in consideration of the simplicity, model two only has 5 variables while model one has 8 variables. Even though model one maybe more accurate and can make more profit. We still choose the model two since it is simple and stable. We could put in into practice easily and guarantee its result.

Screening Tool

To start with, every variable is mapped into questions and assigned weights for each based on the z value. As we can see below, the weight is not a simple round percentage, so they would be rounded a little bit to make the score calculation more straightforward. We initially put a 45% weight on age, 15% each on hypertension, CVD, and diabetes, then the final 10% for white race. Age is divided into three different categories: <40, 40-64, >64. Any age below 40 is given no point (0/2*45%), age between 40 and 64 is given point of (1/2*45%), and any age above 64 is given the full point of 45%. The other variables are treated as a binary. We added a couple races other than white and assign zero points for them to reduce potential misunderstanding of racism if only using "white or others" category. The reason why 64 is chosen as the cutoff point is because the density of the data with people over age 64 that has CKD. 375 out of 464 (80.82%) of the training data who have CKD are people age over 64, indicating a very huge proportion

Using that model on the outside sample that has missing values imputed using mice package, it is found that there are 375 out of 2819 that has inconsistent result between R calculation and survey calculation. Furthermore, the weight on age alone is too high on the survey. If there is a younger white person with hypertension, CVD, and diabetes, they might be diagnosed as a non-CKD potential, thus increasing the chance for false negative (not costing us anything in this assignment but could potentially lead to expensive lawsuit in real-world application). Applied to a white older person that has none of the above diseases, that person could be diagnosed as a CKD potential merely because he is old, causing a false positive signal.

Variable	*Z-value*	*Weight*
Racegrpwhite	3.862	0.1041588
Age	18.331	0.4943902
Hypertension	4.719	0.12727224
CVD	4.929	0.13293597
Diabetes	5.237	0.14124279

With that in mind, each variables age, hypertension, CVD, diabetes, and race white are now re-weighted to 30%, 20%, 20%, 20%, and 10% respectively to be more equally representative and to improve alignment with the R model. With that model, the alignment improved to only 268 on difference between survey model.

Limitations

1) The study population is not a random sample of U.S adults, so our model built on the dataset can’t be applied directly to the U.S population and shouldn't be used for actual decision-making.

2) There are 6000 observations within which only 4136 observations are without any missing data. As a training dataset to predict if people have CKD, 6000 observations are far not enough to guarantee an accurate model.