SCREENING FOR CHRONIC KIDNEY DISEASE REPORT

EUN JU JONG

EXECUTIVE SUMMARY

Chronic Kidney Disease (CKD) is a condition characterized by a gradual loss of kidney function over time, which is associated with high risk of cardiovascular disease, kidney failure, and other complications. CKD in primary care is largely asymptomatic and the pathological condition in which CKD is developed is often unknown. Studies have shown that diabetes and hypertension are the two main causes of CKD which are responsible for up to two third of its causes. Strategies for early identification and treatment of individuals with chronic kidney disease can help prevent the progression of kidney disease to the end stage of kidney failure.

This report discusses a possible screening tool for early identification of such high-risk individuals to contribute to reducing morbidity and mortality associated with CKD. We analyzed demographic and health information data collected from 1999 to 2000 and 2001 to 2002 consisting of 6000 individuals who have been reported with and without CKD to identify clinical or sociodemographic factors which largely contribute to detecting individuals with CKD. We combined the evidence from the existing literature with our data analysis and conclude that age, hypertension, diabetes, weight, body mass index (BMI), and history of cardiovascular disease are the major factors that increase the risk of CKD. In addition, high risk group includes those who are female and who have family history of CKD.

The current analysis also provides shortcomings of the analytical techniques used to produce the predicted outcomes to open up possibilities for improvement in CKD detection in the future.

WHAT ARE THE MAIN CAUSES OF CKD?

The existing medical literature suggests that diabetes, hypertension, and age 60 or above are the most critical factors that increase the risk of CKD. It is also suggested that cardiovascular disease, family history of CKD, and ethnic and racial minorities are considerable risk conditions for CKD.

We explore the sample data of 6000 individuals with and without CKD provided by the National Center for Health Statistics of the Centers for Disease Control and Prevention (NCHS-CDC). Exhibit 1 displays the distribution of individuals with and without hypertension from the sample individuals with and without CKD. We see that individuals who have both hypertension and CKD are about 80 percent of the CKD population while individuals who have CKD with no hypertension are about 21 percent of the CKD population.

Exhibit 1

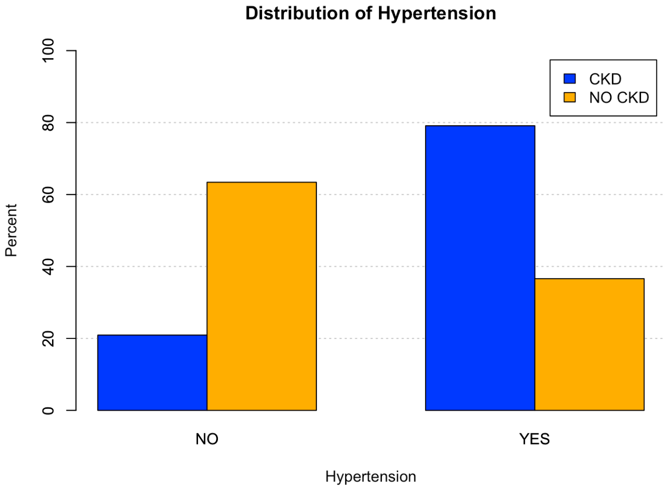


Exhibit 2

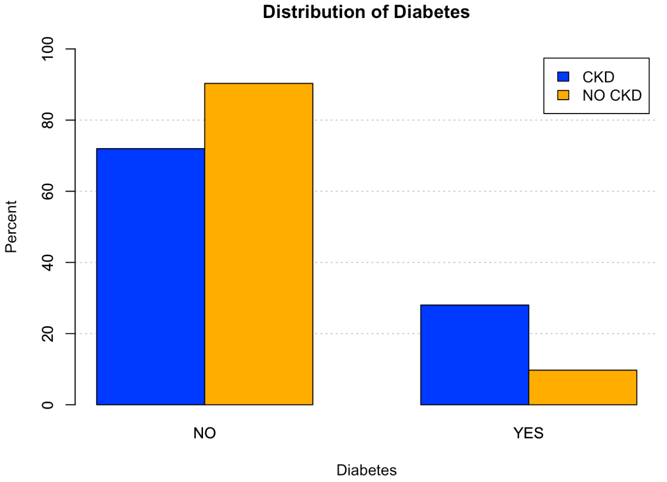
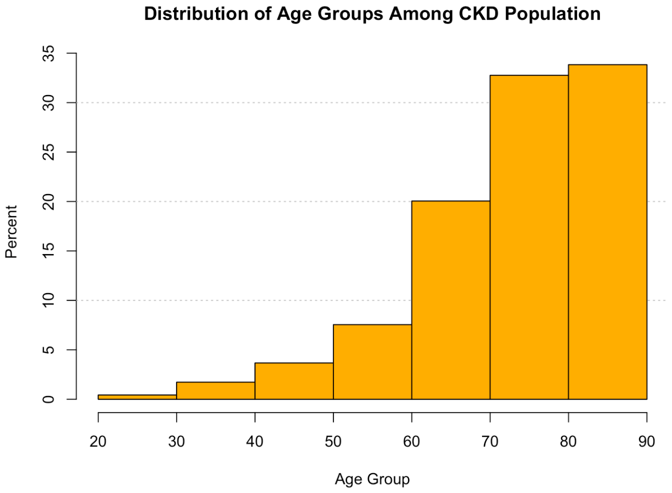


Exhibit 2 displays the distribution of individuals with and without diabetes from the sample individuals with and without CKD. We see that about 28 percent of the CKD patients has reported being diabetic while about 72 percent of the CKD population has reported being diabetic.

Exhibit 3 displays the distribution of age groups of 464 patients who reportedly have CKD. We see that greater percentages of older patients have CKD. Especially, the percentage of CKD patients increases from less than 10 percent to more than 20 percent as the age group jumps from the 50s to the 60s. The increase in the percentage of CKD patients between the age groups of the 60s to 70s and 80s is also drastic with an increase of about 13 percent.

From our sample data, we identify that hypertension and diabetes occur at higher percentages for individuals who have CKD relative to individuals who do not have CKD. In addition, the current data also show that CKD is more common among older people, especially those who are 60 years old or above. To challenge the traditional view on the risk factors of CKD, we examine all the possible risk factors provided by the NCHS-CDC and identify factors that are most likely to predict individuals at high risk of CKD.

Exhibit 3



DATA REDUCTION TECHNIQUES

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) extracts the important information from a multivariate dataset and expresses this information as a set of few new variables which correspond to a linear combination of the original variables. We utilize this data reduction technique to narrow down the number of potential factors to be used in the current analysis. Our goal here is to determine the most contributing factors among the entire variables.

Exhibit 4

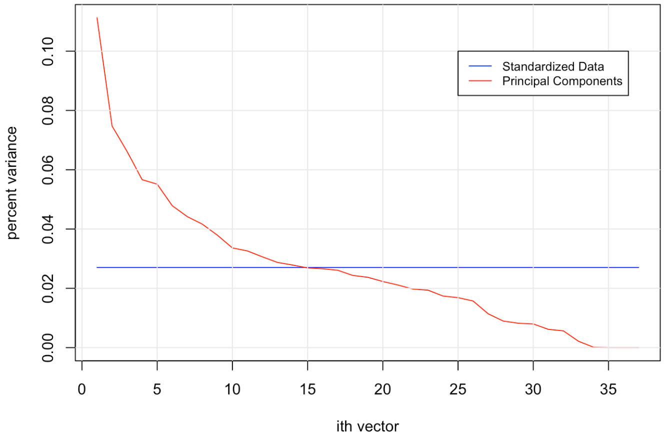
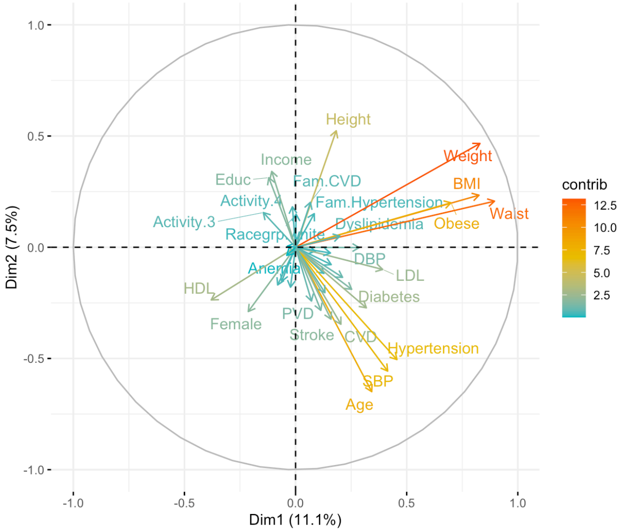


Exhibit 5



We looked at the percent variance of the standardized variables and the principal components presented in Exhibit 4. The original variables have an equal variance due to standardization. From the variances of the principal components, we see that the first principal component accounts for about 11 percent of the variance in the data, and the second principal component accounts for about 7.5 percent of the variance in the data.

Exhibit 5 represents the contribution of each of the variables to principal components 1 and 2. Note that the color grid represents the contribution of variables in accounting for the variability in principal components 1 and 2 are expressed in percentage. We see that a cluster consisting of weight, waist, BMI, and obesity have the highest contribution to principal components 1 and 2. In addition, age, systolic blood pressure (SBP), and hypertension also form another cluster which has a moderately high contribution to principal components 1 and 2. It is suggested that principal component 1 explains about 11.1 percent of the variance in the data while principal component 2 explains about 7.5 percent of the variance in the data.

By looking at the information of the most important variables for principal components 1 and 2, two possible assumptions could be made regarding individuals who are at risk of CKD:

1) Individuals who are overweighted relative to their height and large in size have higher probabilities of getting CKD.

2) Individuals who are old and who have high blood pressure leading to the symptom of hypertension have higher probabilities of getting CKD.

These two assumptions are also associated with the two preliminary causes of CKD suggested by the existing literature. That is, overweight is a risk factor for diabetes, and age and high blood pressure are risk factors for hypertension. Therefore, individuals who carry such conditions would be at higher risk of getting CKD.

THE MOST SIGNIFICANT INDICATORS OF CKD

LOGISTIC REGRESSION

According to the existing literature on the risk factors of CKD, it is suggested that age, hypertension, SBP, diabetes, obesity, history of cardiovascular disease, bad cholesterol (LDL), and family history of CKD potentially increase the risk of CKD (Hallan et. al (2006); Sharma et. al (2013); Vassalotti et. al (2010); Zhang et. al (2007)). Based on the PCA on the current sample data, we have identified that the major contributing factors to CKD are weight, waist, BMI, obesity, age, SBP, and hypertension.

In order to examine the significance of additional sociodemographic factors, we ran a logistic regression using backward elimination on a training dataset and found that female is a statistically significant factor in determining CKD.

We then evaluated the statistical significance of the combination of risk factors suggested from the literature, the major contributing factors identified from the PCA, and the gender factor suggested by the general logistic model.

Among these variables, we identified that the most significant variables in predicting CKD are as follows: Age, Female, Hypertension, Diabetes, History of Cardiovascular Disease (CVD), Weight, and BMI.

Exhibit 6

	Age	Female	Hypertension	Diabetes	SBP	Weight	BMI	CVD	Predicted Probability	CKD
167	24	Y	Y	N	118	66.6	25.38	N	0.2%	N
842	67	N	N	N	122	73.8	27.01	N	4.8%	N
1565	23	N	N	N	124	76.1	25.9	N	0.08%	N
2499	44	Y	N	N	116	69.6	25.91	N	0.9%	N
3869	72	Y	Y	N	129	94.4	36.46	N	20.4%	N
988	71	N	Y	N	143	96.4	33.87	N	14.3%	Y
2095	81	Y	Y	N	155	76.8	31.93	N	32.3%	Y
3368	85	Y	Y	N	182	62.9	26.8	N	39.7%	Y
3703	73	N	Y	N	122	81.3	26.98	N	11.0%	Y
5229	83	Y	Y	N	174	61.4	26.86	Y	52.6%	Y

We utilized a kNN Imputation method, which finds the k nearest samples via Euclidian distance and imputes the mean of those samples, to fill in missing data for our dataset.

We computed a logistic regression with our best factors to estimate the probability of having CKD and tested the model on the training dataset. Exhibit 6 displays the predicted probabilities of 5 individuals who have been reported with CKD and 5 individuals who have been reported with no CKD.

From Exhibit 6, we see that the predicted probabilities of individuals with CKD range from 11 percent and above while the predicted probability of individuals with no CKD mostly lower than the 1 percent level. However, we notice that the individual with ID number 3869 has a relatively high predicted probability of 20.4% compared to the other individuals with no CKD. From the profile of this individual, we can identify multiple risk factors of CKD such that this individual is a 72-year-old woman who has hypertension and who is overweight considering that her BMI falls within the obesity range. Therefore, based on her profile, she is considered to be in the high-risk group of getting CKD even if she does not have CKD at the moment.

MODEL EVALUATION:

PERFORMANCE DIAGNOSTIC TESTS

It is suggested by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) that the prevalence of CKD among the U.S. general population is approximately 14 percent. We utilized 14 percent prevalence of CKD as a threshold to classify individuals into positive CKD group and negative CKD group based on the predicted probability of the sample population with or without CKD.

Exhibit 7 displays the total of predicted outcomes and actual classification of CKD patients at 14 and 30 percent thresholds for comparison. We consider the model with the 14 percent classification threshold. In a total of 5071 (346+4725) cases out of 6000 cases, the current model accurately predicted that individuals either have or do not have CKD, which marks the accuracy of the model to be about 84.52 percent. On the other hand, in 811 cases out of 6000 cases, the model predicted that individuals had CKD while they did not have CKD, and in 118 cases out of 6000 cases, the model predicted that individuals did not have CKD while they actually had CKD. The accuracy of the 14 percent threshold is about 84.52 percent while the accuracy of the 30 percent threshold is about 91 percent.

RISK OF MEDICAL TEST "FALSE NEGATIVE"

Exhibit 8 displays an ROC Curve which represents the performance of all classification thresholds. The curve can also indicate the tradeoffs between sensitivity and specificity of the CKD classification of the current model. Note that sensitivity measures the ability of the model to correctly generate a positive result for people who have the condition that is being tested for, and specificity measures the ability of the model to correctly generate a negative result for people who do not have the condition that is being tested for.

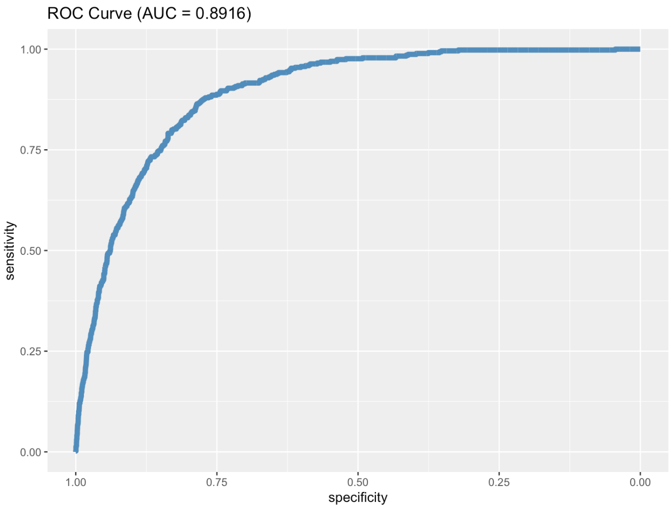
On the x-axis is specificity, and on the y-axis is sensitivity. An increase in specificity leads to lower sensitivity and vice versa. AUC represents the area under the ROC curve, which is a measure of the classification performance of the current model. The overall accuracy of the model is about 89.16 percent.

Consider that our classification is medical diagnosis, we need to evaluate whether the overall diagnosis ability of model is good enough. One way is to look at the accuracy of the model. However, beside the accuracy of the model, it is also important to the outcomes of false positive and false negative diagnosis.

Exhibit 7

14% Threshold	Referenced	
Predicted	CKD	NO CKD
CKD	346	811
NO CKD	118	4725
30% Threshold	**Referenced**	
Predicted	CKD	NO CKD
CKD	199	273
NO CKD	265	5263

Exhibit 8



The danger of false negative is that it gives both the patient and the doctor who perform the diagnostic test a false sense of security that the patient does not need medical treatment. This essentially leads to failure of the patient to receive an appropriate treatment in an early stage.

Based on the performance matrix in Exhibit 9 for the 30 percent classification threshold, the sensitivity of the model decreases from about 74.57 percent to about 42.89 percent. By increasing the sensitivity of the model, the problem occurs with the risk of false negatives being about 57 percent. The tradeoff between an increase of about 6.5 percent of accuracy as well as less cost for false positive diagnosis and the increased false negative diagnosis should be considered in this case. From a perspective of medical diagnosis, we would prefer to keep the probability of false negative to be low. Therefore, a 14 percent threshold is selected for our predictive model.

Exhibit 9

14% Threshold			
Accuracy	84.52%	Sensitivity	74.57%
Precision	29.9%	Specificity	85.35%
30% Threshold			
Accuracy	91%	Sensitivity	42.89%
Precision	42.19%	Specificity	95.07%

CONCLUSION

The current analysis utilized the data reduction technique of PCA to identify factors which have the most contribution to the characteristics that represents individuals who have high probability of getting CKD. This technique allows us to unveil that individuals who are overweighed and who have high BMI are at higher risk of getting CKD relative to individuals who are not overweighted and whose BMI are in the healthy range. In addition, individuals who are old and who have high blood pressure, especially systolic blood pressure (SBP), leading to the symptom of hypertension are more likely to get CKD relative to individuals who are younger and who do not have high blood pressure.

Using the logistic regression analysis, we identified that the main risk factors of CKD are Age, Female, Hypertension, Diabetes, History of Cardiovascular Disease (CVD), Weight, and BMI. We applied the 14 percent prevalence of CKD in the U.S. general population as the threshold for CKD classification of our model. The results show that the current model acquires approximately 89.16 percent of accuracy in predicting whether an individual has CKD. With the current classification threshold, the model maintains about 85.35 percent of specificity and about 74.57 percent of sensitivity. The use classification threshold at a 14-percent level produces about 25 percent of false negatives in CKD classification. The cost of the decision of the 14 percent threshold is the tradeoff the between less false negative detection and more false positive detection which would cause additional medical expenses to the patients who are falsely predicted with CKD in the real-world setting.

Limitations

The current analysis utilized the data collected in as early as 1999. Though it is unlikely that the clinical aspects the variables are considered out of date for such a small-scale analysis, it is possible that the relevant factors to CKD may have changed overtime. Therefore, acquiring more recent data would allow us to gather more relevant information and conduct up to date analysis.

The current study used the 14 percent threshold based on the prevalence of CKD in the U.S. general population with the cost of less accuracy of the model and additional costs occurred to people who are predicted to have CKD while they do not. This decision was based on reducing false negative detections to a reasonably low level, so we did not target the threshold that produced the greatest accuracy for detecting CKD.

Logistic regression is a statistical analysis model that is used to predict probabilistic outcomes based on independent factors. So, there is a risk of overfitting as well as overstating the accuracy of predictions when we try to fit a high-dimensional dataset.

Hypertension and diabetes are the two main well-known risk factors of CKD. If we look at the factors which we identified to be major contributors to CKD in the current model, the majority of these factors are highly intercorrelated with hypertension and diabetes, which introduced the risk of multicollinearity, and this issue cannot be resolved in logistic regression.

References

Facts About Chronic Kidney Disease. National Kidney Foundation. (2021, February 25). https://www.kidney.org/atoz/content/about-chronic-kidney-disease.

Grover, K. (2020, June 23). *Advantages and Disadvantages of Logistic Regression*. OpenGenus IQ: Learn Computer Science. https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/.

Hallan, S. I., Dahl, K., Oien, C. M., Grootendorst, D. C., Aasberg, A., Holmen, J., & Dekker, F. W. (2006). Screening strategies for chronic kidney disease in the general population: follow-up of cross sectional health survey. *Bmj*, *333*(7577), 1047.

James, M. T., Hemmelgarn, B. R., & Tonelli, M. (2010). Early recognition and prevention of chronic kidney disease. *The Lancet*, *375*(9722), 1296-1309.

Manns, B., Hemmelgarn, B., Tonelli, M., Au, F., Chiasson, T. C., Dong, J., & Klarenbach, S. (2010). Population based screening for chronic kidney disease: cost effectiveness study. *Bmj*, *341*.

Orantes, C. M., Herrera, R., Almaguer, M., Brizuela, E. G., Hernández, C. E., Bayarre, H., ... & Castro, B. E. (2011). Chronic kidney disease and associated risk factors in the Bajo Lempa region of El Salvador: Nefrolempa study, 2009. *MEDICC review*, *13*, 14-22.

Sharma, S. K., Dhakal, S., Thapa, L., Ghimire, A., Tamrakar, R., Chaudhary, S., ... & Lamsal, M. (2013). Community-based screening for chronic kidney disease, hypertension and diabetes in Dharan.

U.S. Department of Health and Human Services. (n.d.). *Kidney Disease Statistics for the United States*. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/health-statistics/kidney-disease#:~:text=The%20overall%20prevalence%20of%20CKD,661%2C000%20Americans%20have%20kidney%20failure.

Vassalotti, J. A., Fox, C. H., & Becker, B. N. (2010). Risk factors and screening for chronic kidney disease. *Advances in chronic kidney disease*, *17*(3), 237-245.

Zhang, L., Zuo, L., Xu, G., Wang, F., Wang, M., Wang, S., ... & Wang, H. (2007). Community-based screening for chronic kidney disease among populations older than 40 years in Beijing. *Nephrology Dialysis Transplantation*, *22*(4), 1093-1099.