SCREENING TOOL FOR CHRONIC KIDNEY DISEASE

Aishwarya Reddy Toom
Deepika Burre
Hemachandar Nagarajan
Pinkesh Nayak
Sasidhar Sirivella
Vandana Lunia

ABSTRACT

Chronic kidney disease (CKD) is a type of kidney disease in which there is gradual loss of kidney function over a period of months to years. Initially there are generally no symptoms; later, symptoms may include leg swelling, feeling tired, vomiting, loss of appetite, and confusion. Complications include an increased risk of heart disease, high blood pressure, bone disease, and anemia. CKD can affect almost every body system. Early recognition and intervention are essential to slowing disease progression, maintaining quality of life, and improving outcomes.

Our study implements logistic regression and develops a model to identify whether a person has CKD. We cited various research papers and consulted various Doctors and implemented logistic regression to measure the model's accuracy. We conducted (a) Correlation Analysis, (b) Lasso Regression and Scientific citation for feature Selection. Logistic regression gave a training accuracy of 80.3 % with a validation accuracy of 79.46%. The study also consists of a simple screening tool which most likely indicates the presence of CKD. This study concludes by using predictive model and the screening tool to predict the risk of CKD in 2819 people.

METHODOLOGY

DATA DESCRPTION: The dataset for the case study consists of responses for specifically designed questionnaire from 8819 individuals, aged 20 years or older taken between 1999-2000 and 2001-2002 during a survey across various states in USA. The dataset is divided into two sets 1. Training set with 6000 observations in which 33 variables along with the status of CKD is provided. 2. Testing set consisting of 2819 observations with same set of variables in which the CKD has to be predicted. *Table1* has all the 33 variables given in our dataset.

Age	Female	Education
Income	CareSource	Insured
Height	BMI	Obese
SBP	DBP	HDL
Total.Chol	Dyslipidemia	PVD
Poor Vision	Smoker	Hypertension
Diabetes	Fam.Diabetes	Stroke
Fam.CVD	CHF	Anemia
Unmarried	LDL	CVD
Weight	Activity	Race group
Waist	Fam.Hypertension	CKD

Table1.

MISSING DATA

Missing data can occur because of nonresponse: no information is provided for one or more items or for a whole unit ("subject"). Some items are more likely to generate a nonresponse than others: for example, items about private subjects such as income.

Our dataset consists of 8819 responses against 33 attributes (8819 x 33) 291027 individual responses are to be recorded. But only 283285 are recorded and 7742 records are missing (which is about 2.6 % of the data set). Four dummy variables have been created for Race group (Black, White, Hispanic and others).

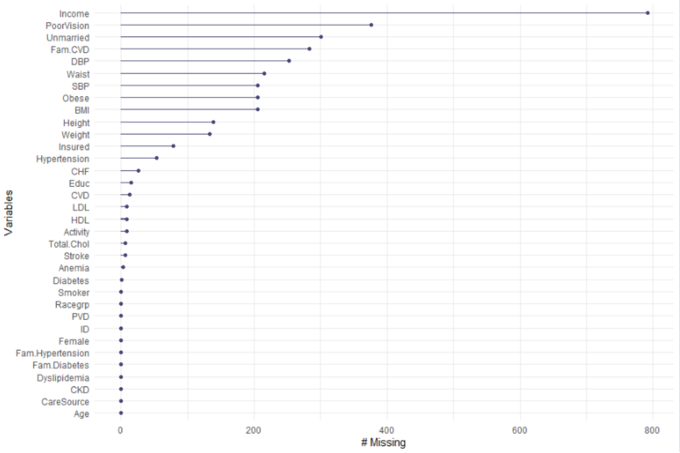


Figure 1

The above picture shows the number of missing values in each of the variables. Further the missing values are analysed to observe any patterns or combinations occur within the data (Picture 2).

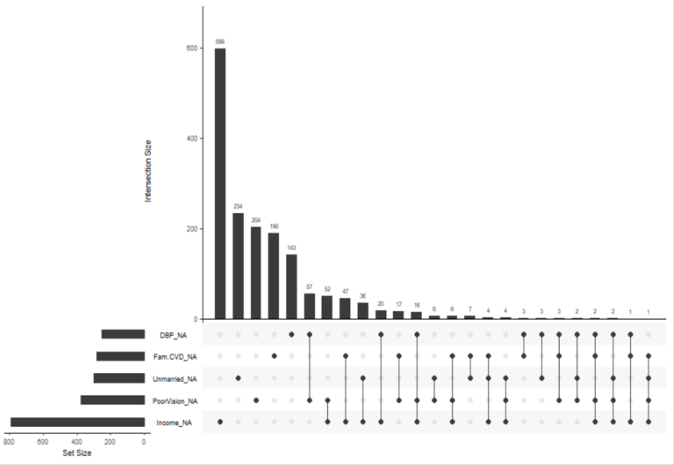


Figure 2

IMPUTATION

Missing data reduces the representativeness of the sample and can therefore distort inferences about the population. The choice of method to impute missing values, largely influences the model's predictive ability. If the data is missing completely at random then deletion does not add any bias, but it might decrease the power of the analysis by decreasing the sample size. To deal with the missing data here, MICE package has been used with mean imputation so that the overall mean will not be affected.

VARIABLE SELECTION

Attribute selection methods are used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. We have used correlation analysis and cited many research papers online and eliminated following variables: Income, Unmarried, CareSource, Insured, Education, Height, Weight, LDL, Total Cholesterol for the initial selection. Then to further filter out the insignificant variables we have used several approaches.

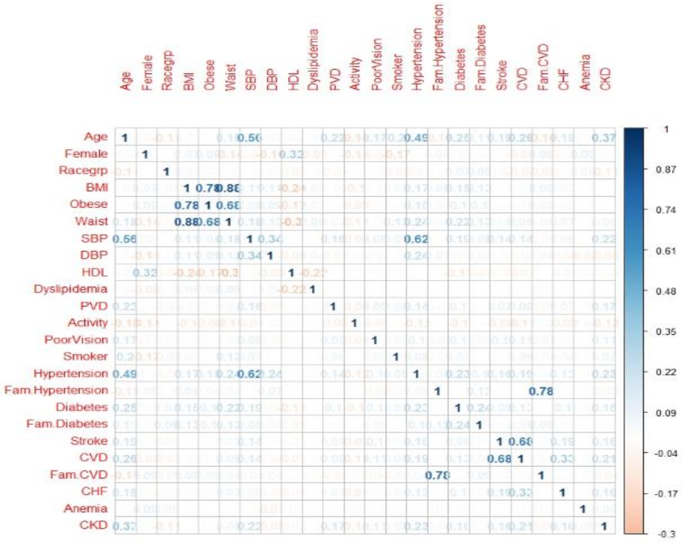


Figure 3

ANALYTICAL APPROACH

We used Lasso regression for feature selection in remaining 24 variables. Based on the lasso model, following 13 variables have higher significance: Age, DBP, HDL, PVD, Activity, Hypertension, Diabetes, Stroke, CVD, Anemia, Racegroup Hispanic.

SCIENTIFIC APPROACH

Considering real life scenario, we cross validated the variables with Nephrologist and modified the above variables and finalized the following 13 variables for our predictive model: Age, Female, BMI, Dyslipidemia, PVD, Hypertension, Diabetes, Family Diabetes, Stroke, Family CVD, CHF, Anemia and Race Group (Black, White, Hispanic and others).

CRITERION BASED APPROACH

The Akaike Information Criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. We want to minimize AIC. Larger models will fit better and so have smaller RSS but use more parameters. Thus, the best choice of model will balance fit with model size.

Different Logit models were run on the different combination of variable selected. We used AIC as an estimator on the all the models and picked out the best one with the lowest AIC value.

PREDICTION MODEL

This logit model gave an AIC value of 1478.8. Then to train and test the model we split the 6000 training set data into three sets 1) Main Training Set (4000) 2) Testing (1000) Set 3) Validation Set (1000).

FRAMEWORK OF THE APPROACH

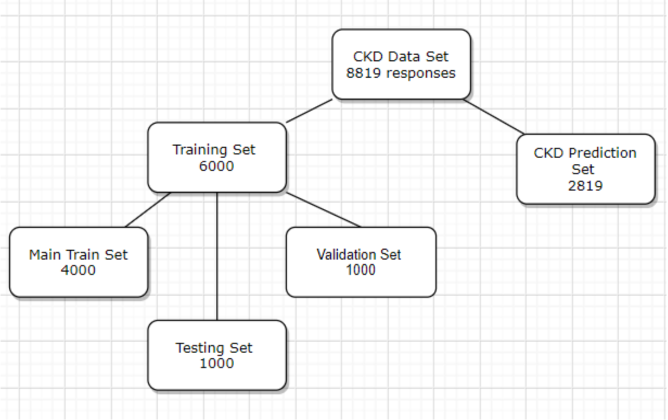


Figure 4

ACCURACY OF THE MODEL

After training the model, we tested it with the Testing set consisting of 1000 responses to check the accuracy of the model and select the threshold for prediction. We generated a for loop on this set to predict CKD for this testing set for various thresholds and compared with the actual CKD values to get confusion matrix, accuracies and corresponding costs.

THRESHOLD SELECTION

We can convert the probabilities to predictions using what's called a threshold value, **t**. If the probability of CKD is greater than this threshold value, **t**, we predict that person has CKD. But if the probability of CKD is less than the threshold value, **t**, then we predict that the person does not have CKD.

How to select the value for t: The threshold value, **t**, is often selected based on which errors are better. This would imply that **t** would be best for no errors but it's rare to have a model that predicts perfectly. In this model we have selected based on money as well. For taking test it takes 100 dollars for FP and TP we would lose 100$ and for TP the hospital will gain 1000$.

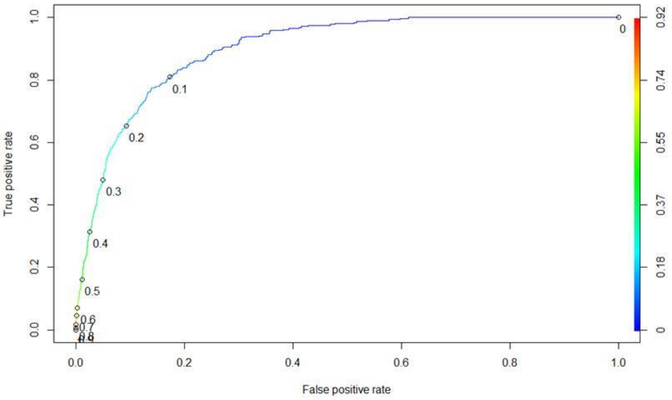
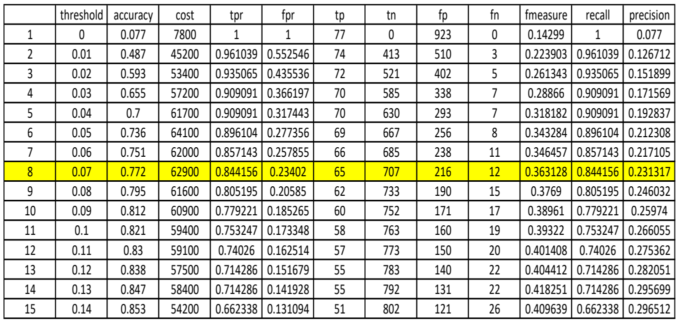


Figure 5

AUC VALUE: We selected the threshold based on the ROC curve. AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, higher the AUC, better the model is at distinguishing between patients with CKD and without CKD. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. The AUC value for the above ROC curve 87 %.



	threshold	accuracy	cost	tpr	fpr	tp	tn	fp	fn	fmeasure	recall	precision
1	0	0.077	7800	1	1	77	0	923	0	0.14299	1	0.077
2	0.01	0.487	45200	0.961039	0.552546	74	413	510	3	0.223903	0.961039	0.126712
3	0.02	0.593	53400	0.935065	0.435536	72	521	402	5	0.261343	0.935065	0.151899
4	0.03	0.655	57200	0.909091	0.366197	70	585	338	7	0.28866	0.909091	0.171569
5	0.04	0.7	61700	0.909091	0.317443	70	630	293	7	0.318182	0.909091	0.192837
6	0.05	0.736	64100	0.896104	0.277356	69	667	256	8	0.343284	0.896104	0.212308
7	0.06	0.751	62000	0.857143	0.257855	66	685	238	11	0.346457	0.857143	0.217105
8	0.07	0.772	62900	0.844156	0.23402	65	707	216	12	0.363128	0.844156	0.231317
9	0.08	0.795	61600	0.805195	0.20585	62	733	190	15	0.3769	0.805195	0.246032
10	0.09	0.812	60900	0.779221	0.185265	60	752	171	17	0.38961	0.779221	0.25974
11	0.1	0.821	59400	0.753247	0.173348	58	763	160	19	0.39322	0.753247	0.266055
12	0.11	0.83	59100	0.74026	0.162514	57	773	150	20	0.401408	0.74026	0.275362
13	0.12	0.838	57500	0.714286	0.151679	55	783	140	22	0.404412	0.714286	0.282051
14	0.13	0.847	58400	0.714286	0.141928	55	792	131	22	0.418251	0.714286	0.295699
15	0.14	0.853	54200	0.662338	0.131094	51	802	121	26	0.409639	0.662338	0.296512

Table 2

Our main aim was to reduce the FP(False Positive), so we went with a threshold which gave us less FP at the same time more profit. That threshold came out to be 0.07 with an accuracy of 77.2%.

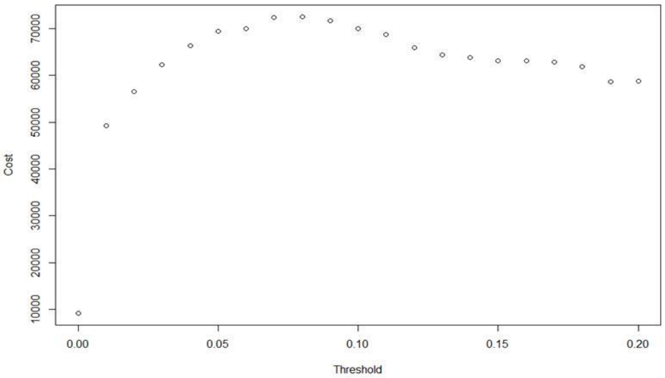
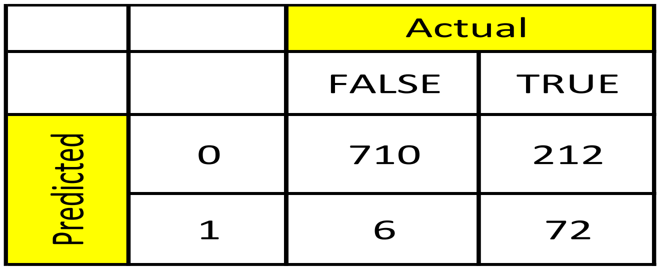


Figure 6

For further validation this logit model was validated with the validation set of 1000. The model gave out an accuracy of 78.2%.



		Actual	
		FALSE	TRUE
Predicted	0	710	212
	1	6	72

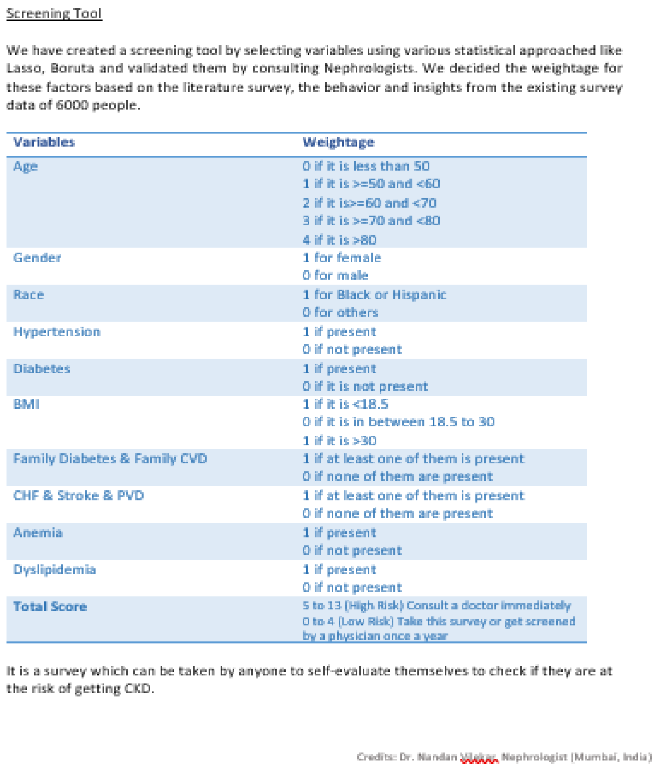
Figure 7

With the above threshold we got a cost of $74,200. Running the model on the Validation set further confirmed our model's accuracy and AUC were almost similar for both the Testing and Validation sets. So, we went with the threshold of 0.07 and finally ran the model to predict whether a person has CKD on the 2819 Prediction Set.

LIMITATIONS OF THE MODEL

- The data set doesn't talk about the severity of CKD (Symptoms like CVD develop only at the final stage. So, if a person who is prone to CKD might not have CVD at the initial stage).
- Data about main causes of CKD are missing (Protein Urea, IHD, eGFR). Two simple tests can detect CKD urine, albumin and serum, creatinine. We don't have these in the data set.
- The variables selected might not be the best combination as more research work is needed on each variable.
- Under-representation of certain race may lead bias in the prediction.
- Imbalance in the data set -We have less people with CKD. This will lead to bias again.
- The data set is focused only for people of US state and during a certain time period. So, putting a general Prediction model is very tough.
- The model was focused on reducing FP and maximizing the profit but the most dangerous one is FN (False Negative).

SCREENING TOOL



Screening Tool

We have created a screening tool by selecting variables using various statistical approached like Lasso, Boruta and validated them by consulting Nephrologists. We decided the weightage for these factors based on the literature survey, the behavior and insights from the existing survey data of 6000 people.

Variables	Weightage
Age	0 if it is less than 50 1 if it is >=50 and <60 2 if it is>=60 and <70 3 if it is >=70 and <80 4 if it is >80
Gender	1 for female 0 for male
Race	1 for Black or Hispanic 0 for others
Hypertension	1 if present 0 if not present
Diabetes	1 if present 0 if it is not present
BMI	1 if it is <18.5 0 if it is in between 18.5 to 30 1 if it is >30
Family Diabetes & Family CVD	1 if at least one of them is present 0 if none of them are present
CHF & Stroke & PVD	1 if at least one of them is present 0 if none of them are present
Anemia	1 if present 0 if not present
Dyslipidemia	1 if present 0 if not present
Total Score	5 to 13 (High Risk) Consult a doctor immediately 0 to 4 (Low Risk) Take this survey or get screened by a physician once a year

It is a survey which can be taken by anyone to self-evaluate themselves to check if they are at the risk of getting CKD.

Credits: Dr. Nandan Vilekar, Nephrologist (Mumbai, India)

Table 3

CONCLUSION

The model can accurately identify patients receiving low-quality care with test set accuracy being equal to 78 % with 13 attributes. In practice, the probabilities returned by the logistic regression model can be used to prioritize patients for intervention. Any individual's risk can be estimated as the probability of that individual with the questions used in the simple screening tool. Further research is needed to simultaneously assess the role of multiple risk factors which were not provided in the case study (as mentioned in the limitation section) to validate this model in other population.

REFERENCES

1. Chronic Kidney Disease: Early Education Intervention by Judy Kauffman, MSN, RN, CNN Charlottesville, Virginia (A DNP Scholarly Project presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Nursing Practice School of Nursing University of Virginia May 2017).
2. Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes by Asif Salekin and John Stankovic (Department of Computer Science University of Virginia Charlottesville, Virginia).
3. An Introduction to Statistical Learning with Applications in R by Gareth James and Daniela Witten.
4. Centre for Disease Control and Prevention (https://www.cdc.gov/kidneydisease/publications-resources/2019-national-facts.html)
5. African Health Sciences https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4915439/)

APPENDIX:

Figure 1 All the variables in the dataset are listed in the table1. Explanation of each variable is given in the below table.

Variable	Data type
Age	Age(years)
Female	1 if Female
Race Group	Reported Ethnic Group (white, black, hispanic and other)
Education	1 if more than high school
Unmarried	1 if unmarried
Income	1 if household income is above median
Caresource	Source of medical care(Dr./HMO, Clinic, noplace and other)
Insured	1 if covered by health insurance
Weight	Weight (kg)
Height	Height(cm)
BMI	Body Mass Index (kg/m^2)
Obese	1 if BMI is greater than 30 kg/m^2
Waist	Waist circumference (cm)
SBP	Systolic blood pressure (max)
DBP	Diastolic blood pressure (min)
HDL	(md/dL) the good cholestrol
LDL	(md/dL) the bad cholestrol
Total Cholestrol	(md/dL) the sum of good and bad cholestrol
Dyslipidemia	Too high LDL or too low HDL
PVD	Peripheral Vascular Disease reflected by reduced SBP at the leg relative to the arm
Activity	(1): stand or walk a lot, (2) lift light loads or climb stairs often , (3): heavy work and heavy loads(4)
Poor Vision	Self reported poor vision
Smoker	Smoked atleast 100 cigarettes
Hypertension	The presence of atleast one of four indicators of high blood pressure
Family Hypertension	Family history of hypertension (high blood pressure)
Diabetes	Self reported physician diagonsed or lab test result
Family Diabetes	Family history of diabetes
Stroke	Self reported response to "Has a doctor ever told you that you had a stroke?"
CVD	Response to "Has a doctor every told you that you angina pectoris, myocardinal infraction or stroke?"
Family CVD	Family history of cardiovascular disease
CHF	Self reported response to "Has a doctor ever told you that you had congestive heart failure?"
Anemia	Treatment for anemia received in past three months or hemoglobin at exam lower than 11g/dL
CKD	Chronnic kidney disease as indicated by measured serum creatinine

Figure 1: Variables with highest missing values are plotted in this graph. Income has largest missing data in the dataset.
Figure 2: Patterns in the missing data is plotted in this graph. For example, Income and poor vision are missing in 52 observations.
Figure 3: Correlation Matrix of the Variables in the data set.
Figure 4: Framework explains about the approach we followed to split the data for training the model.
Figure 5: ROC explains the pattern of sensitivity and specificity for different threshold values for our data.
Figure 6: Graph explains cost vs threshold for our data. Maximum profit we get for our predictions at threshold 0.6.
Table 2: Weightage for each attribute in the screening tool is explained in this table.
Figure 7: Confusion Matrix for the Predictions in the Validation Set.