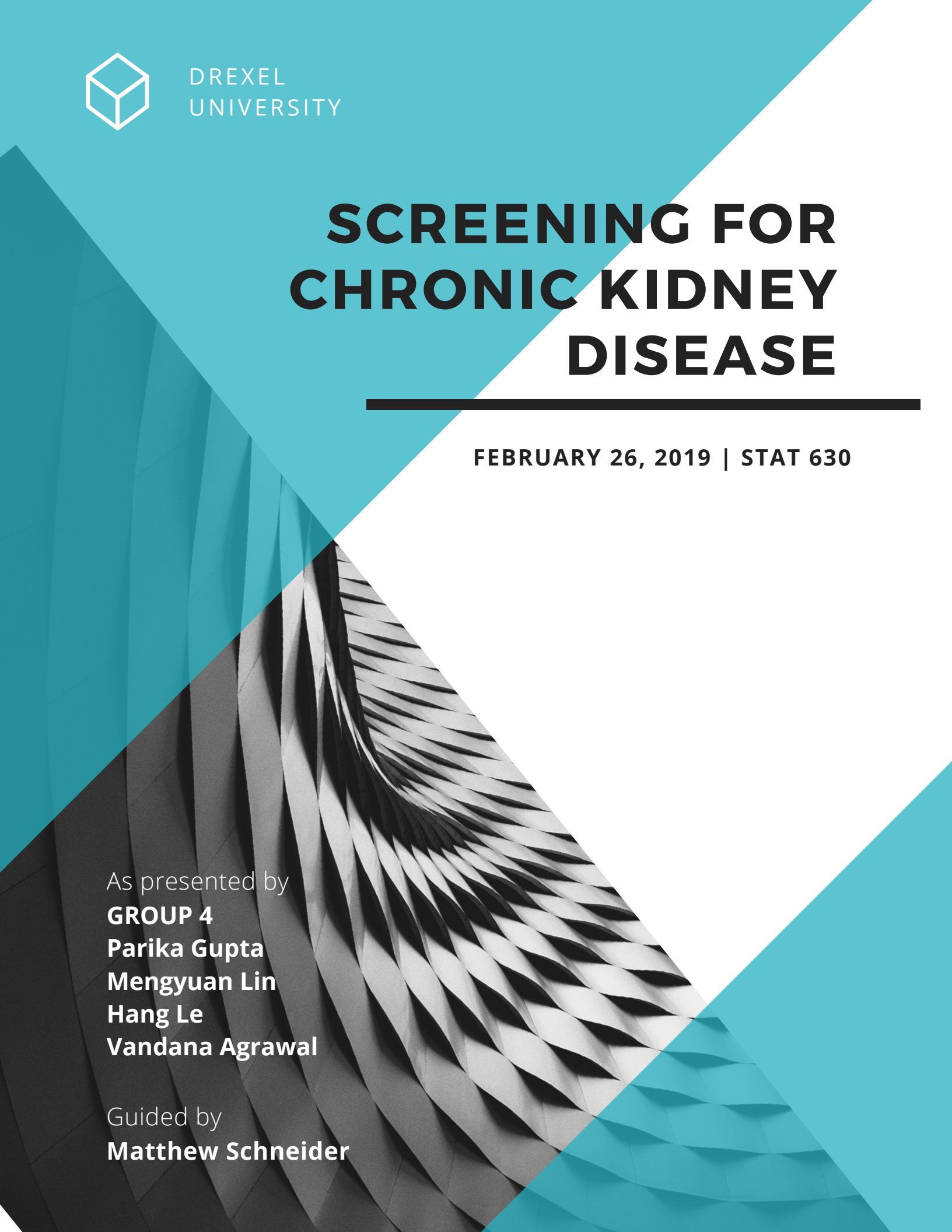
DREXEL
UNIVERSITY

SCREENING FOR CHRONIC KIDNEY DISEASE

FEBRUARY 26, 2019 | STAT 630



As presented by
GROUP 4
Parika Gupta
Mengyuan Lin
Hang Le
Vandana Agrawal

Guided by
Matthew Schneider

INTRODUCTION

CKD, a condition recognized by a gradual loss of kidney function over a period of time, has been receiving growing public awareness for its impact on not only a person's health status, but impaired life quality, shortened life expectancy, as well as excessive healthcare expenditures. Each year, kidney disease has killed more people than most of other diseases for its non-obvious and undetected symptoms in the early stage. A simple-to-use screening tool is in desperate need to identify people who are in high risk of having CKD, and hopefully, the treatment could be applied at the soonest as possible..

OBJECTIVE

In this study, we are to detect at-risk patients from a previous obtained clinical dataset, to define a series of factors that could indicate the presence of CKD, and to develop our credible and easy-to-use screening tool to measure an overall risk for random individual. The dataset contains various information-including demographic (age, gender, and race group) as well as physical record (whether an individual has a certain kind of disease in the past or not) of 8,819 individuals.

An initial analysis on the dataset has generated multiple pre-assumptions for the study. Figure 1 gives the mean comparison on the numeric information, such as age, blood pressure level, and HDL level, between the population carrying CKD and the population free from the threat of CKD. (CKD=1 indicates the CKD carrier group, CKD=0 indicates the group not having CKD)

People at higher ages tend to have higher risk of having CKD. Compared at an average age of 73 for CKD carrier group with an average age of 47 for the other group. Also, a higher SBP, which indicates if a person has high blood pressure, were detected among population that has CKD.

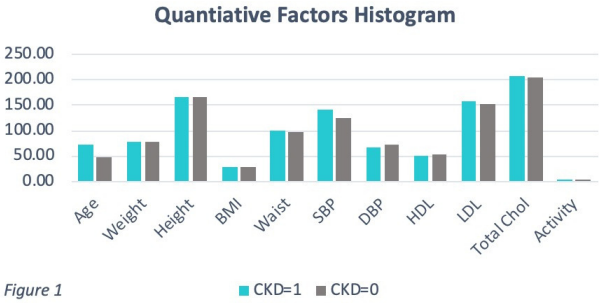


Figure 1

Age and high blood pressure, as a result, were presented as two of the major risk factors that cause the CKD. Figure 2 gives the proportion of population in each category for both groups.

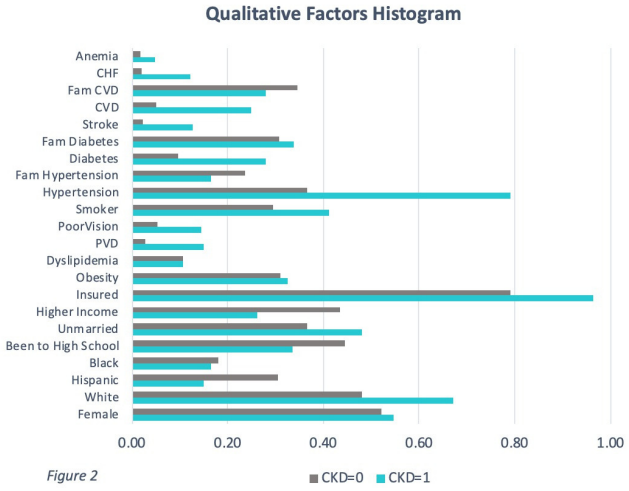


Figure 2

For CKD carrier group, around 79% of the population have Hypertension and over 55% of the population are Female. The comparison between these two groups also shows that there is a larger proportion of the population in no-CKD group has finished high school (45%) and generates higher income (43%) than the CKD carrier group, at 33% and 26%. People that has higher education level and making higher income tend to be more well-informed of the impact of the disease, thus, live healthier lives. This, again, reveal the importance of developing a practical screening tool to assist studied group indicate the presence of CKD.

However, the percentage in Figure 2 is calculated without missing values. The missing values exist for various reasons and would, probably, led us to a false conclusion. Method of handling the missing data is discussing in the following section.

MISSING VALUES

Missing values are frequently encountered while collecting data. The presence of missing values reduces the data available to be analyzed. Missing values also compromise the reliability of the results and cause bias in the analysis which leads to reducing the efficiency of the data analysis. Missing values can arise from information loss or because of dropouts and non-responses of the study participants. Inappropriate handling of missing data can lead to a poor model performance at the model development stage.
The case study for Chronic Kidney Disease has 34 Variables. Definition for each variable is given in Appendix 1. Percentage of the missing data in different categories is presented in Figure 3

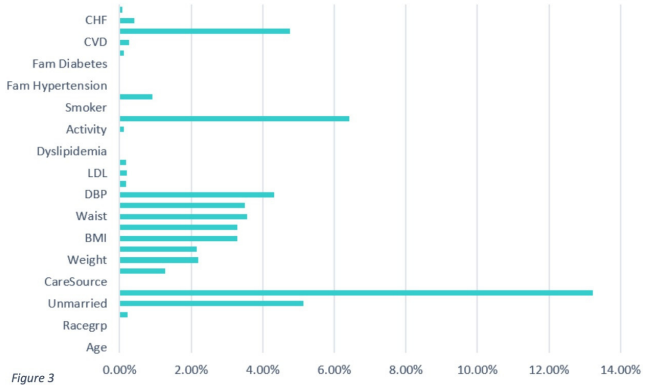


Figure 3

There are multiple ways to handle missing values in a dataset. Below are some ways that can be done including the method to be chosen.
1. Delete anything is missing: All variables that have the most missing values are deleted. The disadvantage of this method is that a large fraction of data gets deleted from the original sample. The results are not efficient to do the analysis.

2. Replacing with average: In this method, the mean of a variable is calculated and replaced with the missing values in their specific column. For example, if the average age of the participants in the study is 68.5 years, then 68.5 is used to replace all missing values for age. This method can bring insignificant and unreliable results in the large datasets.

3. Replace with a random sample: In this method, the missing value is replaced by a random sample for that variable. For example, if age is considered, to replace the missing value, a random sample is taken, and we get 65, then 65 is used to replace that missing value. The disadvantage for replacing with a random sample is that, it results in bias. Another con is that, it is difficult to implement as it is time consuming. The sample has to be run a multiple time to replace every missing value.

4. Multiple imputation: This method is an alternative to overcome the disadvantage for other methods. In this method, each missing value is substituted for a reasonable guess and then analysis is done as a complete dataset. Imputation maintains efficiency of the data analysis.

> "The idea of imputation is both seductive and dangerous"
>
> -R.J.A Little & D.B. Rubin

For our case study dataset, we chose multiple imputation method. The aim of the imputation step is to fill in missing values multiple times using the information contained in the original data sample. It replaces each missing item with a more acceptable value from a set of plausible values, representing a distribution of possibilities.

We selected a package from R to do the multiple imputation. The package is MICE (Multivariate Imputation by Chained Equations). By default, linear regression is used to predict continuous missing values. Once this cycle is complete, multiple datasets are generated. These datasets differ only in imputed missing values. (The code for imputing data can be found in Appendix 2)

For example: Suppose we have X1, X2....Xk variables. If X1 has missing values, then it will be regressed on other variables X2 to Xk. It will consider the other variables as independent variables to predict the values for X1. The missing values in X1 will be then replaced by predictive values obtained using a complete() function.

MODEL FITTING

Once the imputation of the data is done, we split the data into training set and test set. Training set is used to fit our model which we will be testing over the testing set. Division of data is done such that 75% of given data is train data and 25% is test data.

Step 1: Working with the perfect set of data.

We have created a logistic regression model using the perfect train data set, (data without any missing value).

	Confidence Interval		p-Value	Significant
	2.50%	97.50%		
Intercept	NA	71.5214431	0.972997	
Age	0.08321839	0.122142718	2.00E-16	✔
Female	0.2637971	1.273666122	0.004141	✔
Racegrphispa	1.147056	0.031739645	0.06491	
Racegrpother	-0.8480923	1.399648732	0.752328	
Racegrpwhite	-0.3878299	0.591162723	0.704154	
Educ	-0.611717	0.154584373	0.245373	
Unmarried	-0.1216768	0.651117369	0.177461	
Income	-0.1523712	0.583308473	0.40491	
CareSourceclinic	100.864	NA	0.988068	
CareSourceDrHMO	-101.0105	NA	0.988287	
CareSourcenoplace	-101.2531	NA	0.988804	
CareSourceother	-100.957	NA	0.988269	
Insured	0.2059188	1.75787448	0.174207	
Weight	-0.02728437	0.128682466	0.348919	
Height	-0.07620145	0.097903126	0.813379	
BMI	-0.3223157	0.170613187	0.546526	
Obese	-0.385786	0.712631437	0.561389	
Waist	-0.04821883	0.013956763	0.277653	
SBP	-0.01744146	0.001591258	0.104512	
DBP	-0.015419	0.013936579	0.912773	
HDL	-0.02488096	0.000610945	0.066509	
LDL	0.000199815	0.007997199	0.10075	
Dyslipidemia	-0.8473581	0.301914183	0.381668	
PVD	-0.08147153	0.879003982	0.209469	
Activity	-0.5249541	-0.05743365	0.018089	✔
PoorVision	-0.5268324	0.523106929	0.973088	
Smoker	-0.2522897	0.445499484	0.581524	
Hypertension	0.3198899	1.230720468	0.000901	✔
Fam.Hypertension	-0.5981774	0.816195651	0.80396	
Diabetes	0.1291438	0.906336185	0.031888	✔
Fam.Diabetes	-0.3237056	0.333417678	0.865279	
Stroke	-0.5141235	1.143661153	0.458702	
CVD	0.03328674	1.284688601	0.034891	✔
Fam.CVD	-0.6853789	0.596934358	0.957523	
CHF	-0.5648175	0.688003375	0.987334	
Anemia	0.2874077	2.348179701	0.040326	✔

Table 1: Model 1: Model with all non-missing data as train set and all thirty-three-variable considered.
model 1=glm(CKD~.,family="binomial",data=train)

From the above table we can see that age, female, activity, hypertension, diabetes, CVD and anemia are significant variable as their p-value is very less (<.05). All the significant variable has the positive coefficient, meaning that increase in the value of each variable will increase the probability of having CKD. Only Activity variable is negatively correlated to CKD.

Step 2: With the Imputed data: While imputing missing data, we have carefully maintained the density curve of each variable. Hence, we created the model with the imputed dataset which analyses the significant variable in model. Using Logistic Regression, we then created Model 2 with all 33 variables and applied backward elimination approach to identify variables which are most significant. These variables then were included in Model 3.

	Confidence Interval		p-Value	Significant
	2.50%	97.50%		
Intercept	-17.5992	-9.88218	3.26E-12	✓
Age	0.080946	0.106963	2E-16	✓
Female	0.210235	0.983075	0.002527	✓
Racegrphispa	-0.5475	0.376535	0.712309	
Racegrpother	-0.86056	1.234106	0.589157	
Racegrpwhite	0.050094	0.806971	0.028572	✓
Unmarried	-0.00625	0.55196	0.054823	
Height	0.021935	0.061655	3.86E-05	✓
Obese	0.076313	0.888695	0.019845	✓
Waist	-0.02788	0.000381	0.057899	
DBP	-0.01776	0.001917	0.11356	
HDL	-0.02663	-0.00845	0.000174	✓
PVD	0.288917	1.066853	0.000585	✓
Activity	-0.40571	-0.01323	0.038152	✓
Hypertension	0.456272	1.101307	2.46E-06	✓
Fam.Hypertension	-1.05816	-0.00182	0.046673	✓
Diabetes	0.305405	0.93561	0.000108	✓
CVD	0.220302	0.90623	0.001204	✓
Fam.CVD	3.7E-05	0.909926	0.044644	✓
CHF	-0.0012	0.922712	0.048119	✓
Anemia	0.678527	2.196559	0.000184	✓

Table 2: Model 3:
Formula: Model 3 = glm(CKD~Age + Female + RaceGrp + Unmarried + Height +Obese + Waist + DBP + HDL + PVD + Activity + Hypertension + Fam.Hypertension + Diabetes + CVD + Fam.CVD)

From the above table we see that some variables such as Age, Female, Race group, Height, Obese, HDL, PVD, Activity, Hypertension, Diabetes, CVD, CHF and Anemia are significant variables with p- value <.05 .We also verified importance of these variables by checking their confidence interval.

Objective of our analysis is to create easy to use screening tool. Although Model 3 has all significant variables, some of these variables such as CHF, PVD, HDL or family history of hypertension can not be obtained by easy-to-use screening tool. Some people might not be aware of these parameters while filling up the data in screening tool. We have further tried to reduce number of variables in model 3 by building new model (Model 4) including 7 significant variables which are Age, Gender, Activity, Hypertension, Diabetes, CVD and Anemia. According to the case study, Race group also plays an important role in increasing risk of having CKD, so we consider Race Group in our reduced model.

	Confidence Interval		p-Value	Significant
	2.50%	97.50%		
Intercept	8.746509	-6.99672	2E-16	✓
Age	0.074999	0.096253	2E-16	✓
Female	-0.16784	0.341426	0.005802	✓
Racegrphispa	-0.79643	0.083278	0.110317	
Racegrpother	-1.23321	0.816808	0.833144	
Racegrpwhite	0.017461	0.758351	0.043698	✓
Activity	-0.56323	-0.17998	0.000157	✓
Hypertension	0.275551	0.882564	0.000199	✓
Diabetes	0.275526	0.879548	0.000165	✓
CVD	0.324483	0.962079	7.12E-05	✓
Anemia	1.001178	2.439154	2.46E-06	✓

Table 3: Model 4: Formula: Model 4 = glm(CKD~Age+ Female +Racegrp +Activity +Hypertension +Diabetes + CVD+ Anemia)

Comparison between four models

As we can see from Table 4, there is a significant drop in AIC from Model2 to Model3 which means that our Model 3 is more significant compared to Model2. The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better. Analyzing the table, we can see the drop-in deviance in model 3. The accuracy of each model is above 90% which is quite a good result.

Parameters	Model1	Model2	Model3	Model4
AIC	1056.7	1671	1650	1698.1
Null Deviance	1515.32	2518.7	2518.7	2499.1
Residual Deviance	982.73	1597.9	1608.9	1676.1
Accuracy	94.10%	92.90%	93.20%	93.00%
Difference (Null-Residual)	532.5943	920.812	909.764	823.092
pchisq Test	1.14E-89	6.00E-170	7.00E-180	2.00E-170

Table 4: Comparison between 4 models

PREDICTION

Our goal for this analysis is to maximize profit to correctly identify more people with CKD while maintaining reasonable accuracy for the screening tool. To make decision on which thresholds level to apply for our classification, we try to plot accuracy and profit with different thresholds level. (Figure 4)

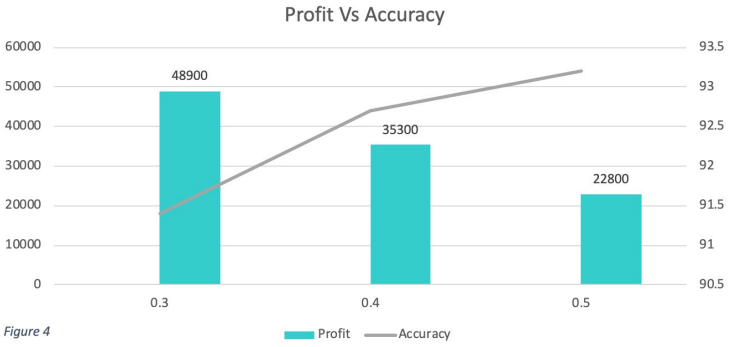


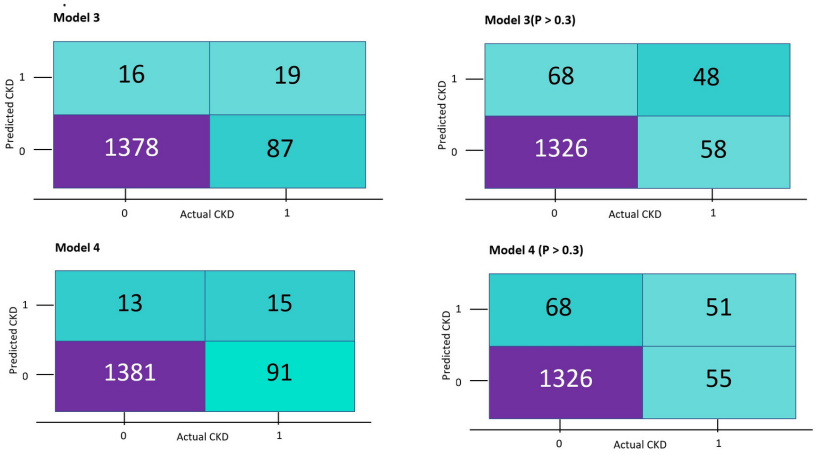
Figure 4

As we can see from the chart, we get the most accurate classification at threshold level 0.5. Our model could predict accurate probability of having CKD for 93% observations in the train data set. However, the cost associated with prediction model will go down drastically due to the fewest number of true positive the model can predict. In contrast, at threshold level 0.3, the cost we can get from prediction model will increase dramatically due to the huge increase in the number of total predicted true positive. However, at this threshold, the model get the lowest level of accuracy, just around 91%, which is still an acceptable accuracy level. Because our goal in this analysis is maximizing the profit, we apply threshold level at 0.3 to classify probability of having CKD among 2819 patients.

As we analyzed in the previous part, two model that we are considering to use for prediction are Model 3 and Model 4. Model 3 presented the most accuracy on the test data set because this model eliminates highly correlated variables in the dataset and only significant variables are kept in this model. In contrast, model 4 is less accurate compared with model 3. The reason we consider taking model 4 is to build simple screening tool to predict chance of having CKD among 2819 patients.

Confusion Matrix

Confusion matrix reflects changes in classification probability from logistic regression result. In Model 3, two confusion matrix were obtained to reflect accuracy level according to change in thresholds. When we decrease threshold to 0.3, there is an increase in both true positive and false positive while there is an huge decrease in number of true negative and false negative. Similarly the number of total true positive and false positive in model 4 also go up according change in thresholds level. We can see from these results below, model 4 only contains 8 variables but this model can predict good results compared to model 3 with 14 variables. This confirms the importance of all 8 variables in model 4 and based on these variables we built our screening tool.



Screening tool

Based on model 4, our screening tool will be developed as a survey including 8 questions to get information about these important variables:

Age: This variable positively correlated with CKD. People who are older than 70 tend to have higher chances of having CKD

Hypertension: Also called as high blood pressure, is one of the leading causes to CKD. Uncontrolled high blood pressure will increase risk of serious health problem including CKD and stroke

Cardiovascular disease (CVD): A class of diseases that involve the heart or blood vessels. If a patient is having CVD, the probability of having CKD will increase.

Diabetes:It is considered as one of the main causes leading to CKD. If a patient has diabetes, the high blood sugar can hurt his kidney over time.

Other important factors: Gender, Race, Anemia and Activity level can increase the chance of having CKD.

Simplicity: The main objective for this screening tool is to be easy to use. Questions that are being asked in our screening tool do not require any extra medical knowledge or exams except one's demographic information and self-health record.

Accuracy: An accuracy check has been implemented on the training dataset. As a result, 88% of the population were correctly classified based on the points generated from our screening tool. The simple screening tool presented less accurate level compared to Model 3 (Table 5). This is because of shortages in some important variable such as HDL and PVD but it's quite impractical to ask in our survey. Appendix 4 shows the difference in prediction between screening tool and our full model.

	Thresholds	Accuracy	True Positive	False Positive	Profit
Model 3	0.5	93.13	19	16	$23,100
	0.4	93.13	34	31	$41,100
	0.3	91.6	48	68	$55,600
Model 4	0.5	93	15	13	$18,200
	0.4	92.5	33	39	$39,000
	0.3	91.1	51	68	$59,500
Screening tool	>60	88.33	58	127	$62,700

Table 5: Comparison between Screening tool and other models

(*) To generate better classification and assign points for each variables for screening tool, detail decision tree will be found in Appendix 5.

LIMITATIONS

Imputation approach was used to replace the missing value for generating a completed dataset for the model prediction. However, the imputing process might potentially weaken the validity of the result and lead us to bias interpretation on it. The significant factors we defined after replacing the missing values might differ from ones where the missing values are present.

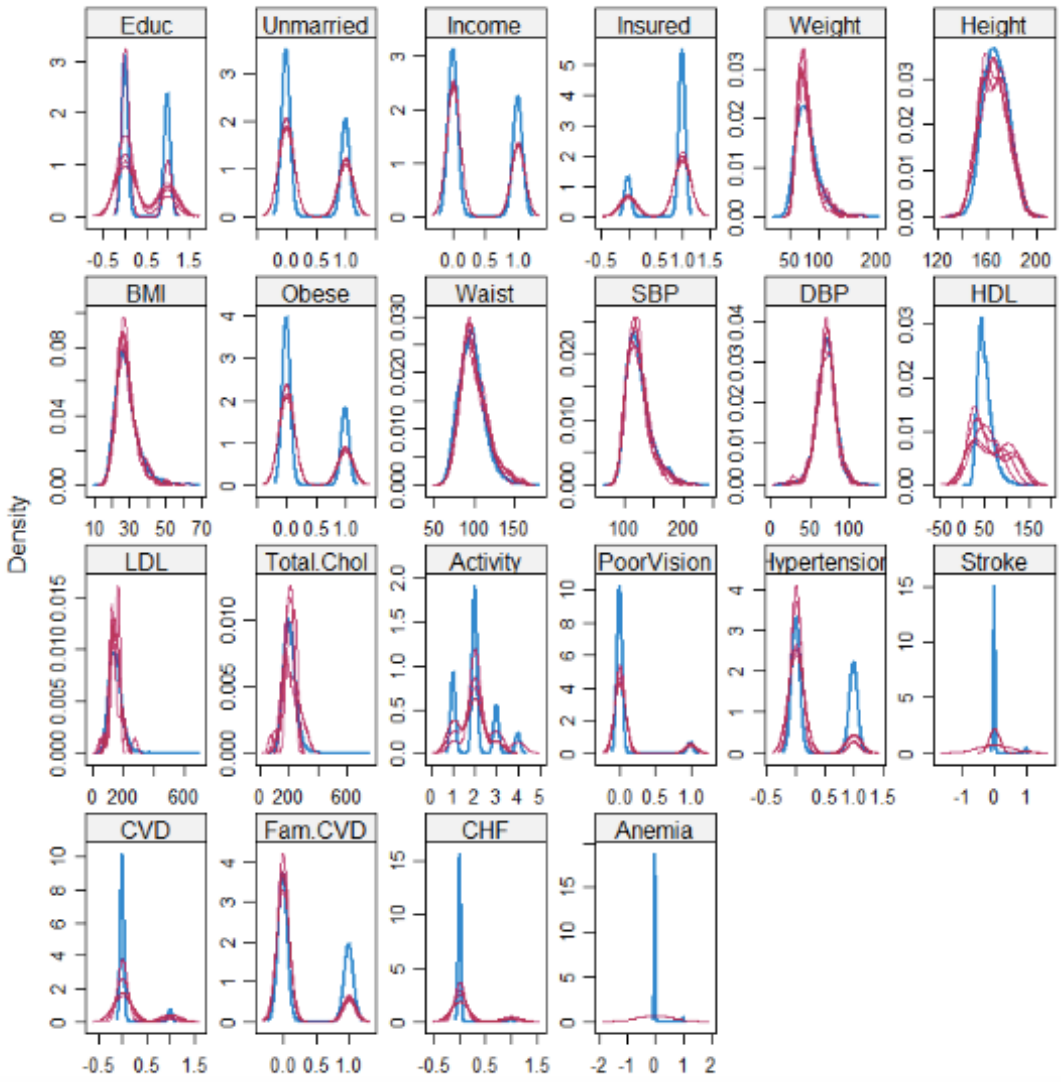
On the other hand, the dataset obtained also limit our ability to detect a perfect set of significant risk factors that put people at high risk of having CKD. For example, family history of kidney disease should be considered but this variable was not included in the dataset.

To overcome the limitations mentioned above, a wider range of data needs to be included. In addition, certain considerable method to obtain data, such as developing clearer and straightforward questions for interviews, should be implemented to minimize the amount of missing values.

APPENDIX 1: DEFINITION OF VARIABLES

Variable	Definition
ID	Identification number
Age	Age (years)
Female	1 if female
Racegrp	Self-reported race/ethnic group (white, black, Hispanic, other)
Educ	1 if more than high school
Unmarried	1 if unmarried
Income	1 if household income is above the median
CareSource	Self-reported source of medical care (Dr./HMO, clinic, noplace, 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	(mg/dL) the "good" cholesterol
LDL	(mg/dL) the "bad" cholesterol
Total Chol	(mg/dL) the sum of good and bad cholesterol
Dyslipidemia	Too high LDL or too low HDL
PVD	Peripheral vascular disease reflected by reduced SBP at the leg relative to the arm
Activity	Mostly sit (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 at least 100 cigarettes
Hypertension	The presence of at least one of four indicators of high blood pressure
Fam Hypertension	Family history of hypertension (high blood pressure)
Diabetes	Self-reported physician diagnosed or lab test result
Fam 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 ever told you that you had angina pectoris, myocardial infarction, or stroke?"
Fam 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	Chronic kidney disease as indicated by measured serum creatinine

APPENDIX 2: DENSITY PLOT OF IMPUTED DATA



APPENDIX 3: IMPUTATION CODES

```
> tempData <- mice(data[1:33],m=5,meth='pmm')
> summary(tempData)
```

- m=5 refers to the number of imputed datasets. Five is the default value.
- meth='pmm' refers to the imputation method. PMM is predictive mean matching and we are using it as imputation method. Other imputation methods can be used, type methods(mice)for a list of the available imputation methods.

Now we can get back the completed dataset using the `complete()` function. With the `complete()` function, we get the complete dataset in which in there is no missing value. The dataset can be used to make models like an original dataset.

```
> data1 <- complete(tempData,1)
> summary(data1)
```

APPENDIX 4: COMPARISON TABLE

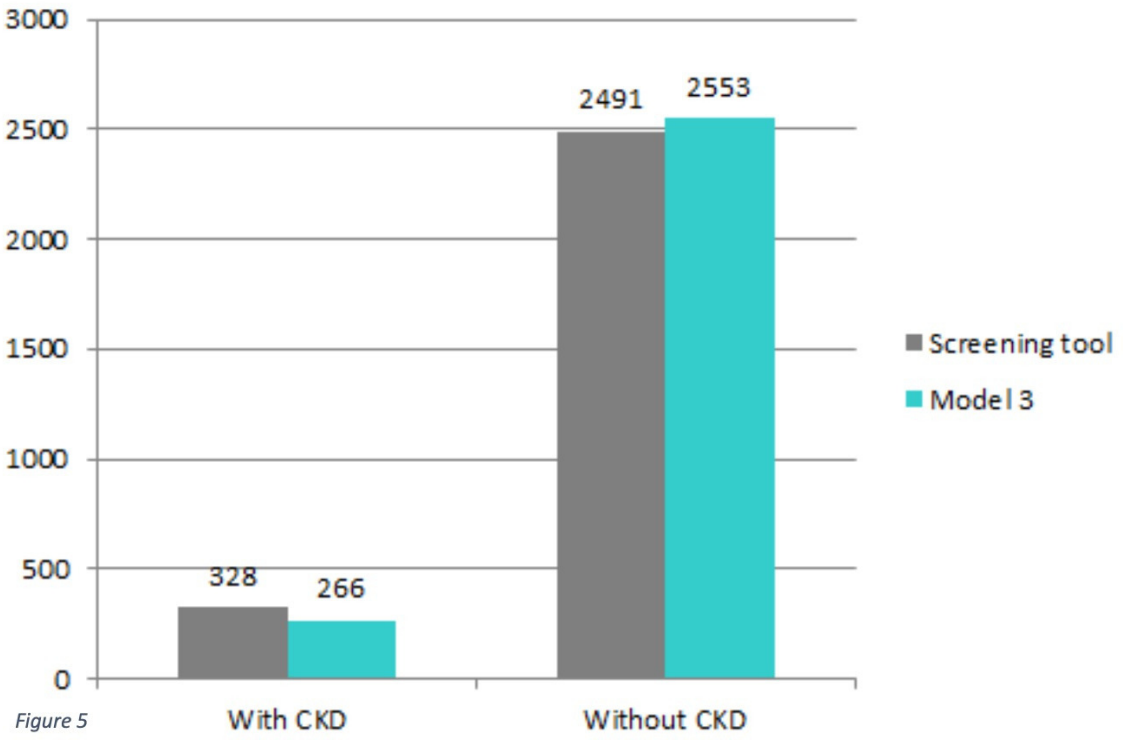


Figure 5

APPENDIX 5: DECISION TREE