DREXEL UNIVERSITY
LEBOW COLLEGE OF BUSINESS
STAT-630 MULTIVARIATE ANALYSIS

# PREDICTIVE ANALYSIS: CHRONIC KIDNEY DISEASE

BUYS, Charlotte
GULATI, Prateek
KHUAT, Trang
WONG, Sze Ki (Zalon)

## EXECUTIVE SUMMARY

The report delivers insights on the prediction model and screening tool developed on observations from 6000 adults in the USA to predict whether people have a high likelihood of developing CKD with an accuracy of 78%. Variable selection was done based on recommendations from medical experts and knowledge gathered from numerous medical journals. The prediction model contained 11 variables out of which 8 were adopted to create the screening tool and scored as per similarity in the range of their odds ratio. Any person who will obtain an 8 or higher score on these variables, as per the scoring system in the questionnaire, is predicted to be at high risk of developing CKD and should undergo further consultation from a medical practitioner. With the dataset provided, our screening tool predicted to generate $61 in profit per person.

## INTRODUCTION

For the last 10 years, the leading causes of deaths were 60% by NCDs, Non-Communicable Diseases.[i] The most common causes are cardiovascular disease, cancer, and Covid-19 for 2020. But looking further down the line, kidney failure takes up the 10th leading cause of death, good for 1.5% of all deaths in the USA in 2020.[ii]

Chronic kidney disease means that there is lasting damage to the kidneys which causes them to not filter blood as they should. Risk factors for developing CKD are diabetes, heart disease, high blood pressure, and kidney failure over the genetic line.[iii] Recognizing kidney failure is complicated since symptoms become recognizable when the kidneys are already badly damaged. That's why testing is important. There are 3 possibilities to test for CKD: eGFR, urine test, and blood pressure.[iv]

In 2018, Medicare beneficiaries who got treatment for CKD had a price tag of over $81.8 billion, or $23,700 per person. The full medical spending for patients with EDSR (or kidney failure) accounts for 7% of the Medicare paid claim costs.[v]

## DATA INTRODUCTION

The dataset was provided by the National Center for Health Statistics of the Center of Disease Control and Prevention and their research was conducted on 8819 adults in the USA. From these subjects, selected information regarding their health was collected, 33 variables in total.

After a simple analysis of the data, missing values were detected. From the training dataset (of 6000 subjects), only 4136 rows have complete data. As for the variables/columns, ID, Female, Racegrp, Smoker, Fam Diabetes, Dyslipidemia, PVD and Fam Hypertension do not have missing values. About 30% of the training set is missing.

After setting up a table with the correlation coefficients of all the variables from the CKD-stats, a few high ones jumped out: LDL & Total Chol. 0,93; Waist & Height 0,88; Waist & BMI 0,88; BMI & Weight 0,86. The fact that these have a high correlation is perfectly obvious since these have a significant relation to each other. Firstly, BMI is calculated by Height and Weight, Obese is measured by Weight and Waist is a good indication of Weight too. And secondly, LDL is one of the measurements for Total Cholesterol.

## DATA PREPROCESSING

The first step we performed in preprocessing data is factorizing the nominal and binomial variables. Since our final imputation method of choice is missForest, factorizing the nominal and binomial variables will ensure that the missForest algorithm does not impute an inappropriate value to our nominal or binomial variables.

As mentioned previously, this data set contains many missing data. In order to handle those missing data, we tested out three different methodologies: complete case, multiple imputations by chained equations (MICE[vi]), missForest; and compare their computational efficiency as well as regression results. Our result suggests that missForest is the most effective imputation method. MissForest is a non-parametric multiple imputation methodology, based on a random forest, developed by Stekhoven and Buhlmann. In missForest, a random forest is trained with the observed values in the first step, then missing values are predicted, and the process repeats iteratively until specified or until the difference between the newly imputed data matrix and the previous one increases for the first time. Theoretically speaking, since the random forest is a considerably stable machine learning model that can handle mixed-type data well, this method appears to be an appropriate imputation method for a data set that potentially has complex interactions and non-linear relations like the data set we have. Error rates from performing the missForest algorithm on our data set suggest that the missForest will impute on numerical variables and categorical variables with 16% and 11% errors, respectively.[vii] Distribution of important numerical and categorical variables in the imputed data set is comparably similar to that in the original data set. (Appendix 1)

As one of the main purposes of building the predictive model to predict CKD is to design a screening tool based on the model, we decide to consider the scoring system and input of our screening tool when pre-processing the data. First, as we decide beforehand to assign scores based on four age bins - below 40, between 40 and 49, between 50 and 59, and above 60 - in the questionnaire, we decide to replicate that scoring system in our model by discretizing Age into the abovementioned 4 bins and factorizing them. Similarly, from our research, we discover that black or African Americans and Hispanic heritages are at a higher risk of CKD than others.[viii] Hence, as we decide to explore these two groups' relationship with CKD in our model building process, dummy variables for these two groups are created.

## MODEL DEVELOPMENT

Adopting the "missForest" methodology for imputation, we recovered a 6000-record data set to build a logistic regression model to design a screening survey for CKD. The data set was split into a training (4001 records) and testing (1199 records) data set based on an 80/20 ratio. Models built on the training data set were then validated on the testing data set and studied for their goodness of fit.

### Selection Metrics

With the goal to create a user-friendly screening survey for CKD in mind, two criteria, namely simplicity and true positive rate (TPR)/false positive rate (FPR), were adopted to select the most suitable model.

Simplicity – We consider a user-friendly survey to be succinct and easy to fill out, in terms of the information requested. We aimed at creating a survey with less than 10 questions, using simple and short sentences, and asking mostly Yes/No questions. Rather than asking about exact medical records like "Do you have a blood pressure above 140/90?", we believe that people who were diagnosed with certain illnesses will be able to correctly answer a question like "Do you have Hypertension?" To summarize, a model involving fewer variables and mostly binary variables were sought.

TPR/ FPR – The goal of the model to correctly predict CKD patients among the sick ones. A high true positive rate (TPR), meaning identifying CKD patients among actual CKD patients, was pursued in the model development process. Also, a low proportion of misclassifying healthy persons as CKD patients (i.e., a low false-positive rate) was preferred. The combination of high TPR and low FPR was concluded by a factor called "profit per person" in the model building process as each true positive classification granted a reward of $1,300 while each false positive case cost $100. The higher the profit per person, the more suitable the model was. To obtain a high profit per person, we understood that the overall accuracy (proportion of correct predictions of both positive and negative cases) was sacrificed.

$$Profit\ per\ person = \frac{1300 * \#True\ Positive\ Case - 100 * \#False\ Positive\ Case}{1199\ (\#record\ in\ test\ data\ set)}$$

Logistic Regression Model Comparison

Several models were built through different methodologies for comparison. A full model utilizing all variables was first developed for the initial study. With variable Age being discretized, the full model contained 38 variables and was studied for multicollinearity. Two sets of variables were identified to be problematic – set 1: HDL, LDL, and Total Cholesterol; set 2: Weight, Height, and BMI. To retain more information and maintain the simplicity of the survey, HDL, LDL, Weight, and Height were considered over Total Cholesterol and BMI in the model building process. Leveraging forward selection and backward elimination, two smaller models were built and used as references to enhance the later models.

Furthermore, we sought some professional advice from a medical doctor for what variables in our data set could be indicative of CKD. After addressing multicollinearity, all suggested variables, including Age, Female, BMI, Obese, Waist, SBP, DBP, HDL, LDL, Dyslipidemia, PVD, Activity, Smoker, Hypertension, Diabetes, Stroke, CVD, and Racegrp, were adopted to construct a model.

Combining the professional advice, the statistical significance of variables shown in the previous models, and the models' performance, we manually selected 18 variables, representing or related to five high-risk factors (age, diabetes, hypertension, cardiovascular diseases, and race) mentioned in the journal article [ix] and built a fusion model. Then, the insignificant variables, whose p-value was higher than 0.05, were removed one by one and tested to determine the profit and other performance metrics.

The performance metrics of the full model, backward elimination, and forward selection models were shown in Table 1 for comparison with the selected models (Doctor's Advice and Fusion Model)

Table 1. Summary of Performance Metrics of Selected models

| | Full Model | Backward Elimination | Forward Selection | **Doctor's Advice** | **Fusion Model** |
|---|---|---|---|---|---|
| *# Variable* | 38 | 21 | 19 | **21** | **15** |
| *AIC* | 1923.3 | 1899.6 | 1903.6 | **1933.2** | **1930.8** |
| *True Positive Rate* | 90.2% | 89.1% | 91.3% | **92.4%** | **90.2%** |
| *False Positive Rate* | 22.0% | 22.0% | 23.8% | **26.3%** | **22.0%** |
| *Profit/Person* | 69.73 | 69.39 | 69.06 | **67.90** | **69.73** |
| *Area under ROC Curve* | 89.4% | 88.9% | 89.1% | **89.1%** | **89.5%** |
| *Accuracy* | 79.0% | 78.8% | 77.3% | **75.1%** | **79.0%** |

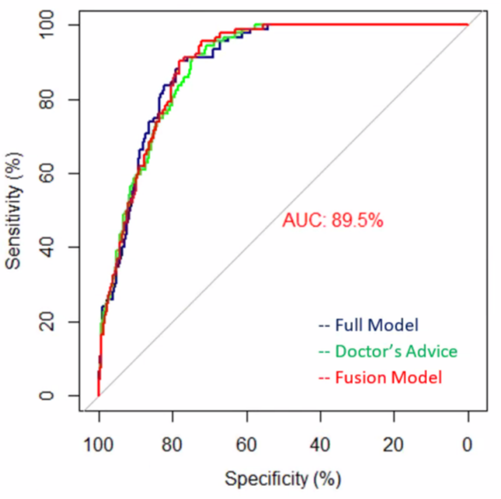
Final Model

The selected models behaved similarly in terms of sensitivity (i.e., TPR) and area under the ROC curve (shown in Figure 1). The fusion model was eventually chosen for its small size (11 original variables) and slightly higher profit per person, which reflected a balance between a high true positive rate and a low false-positive rate.

All 15 variables, except CHF (Congestive Heart Failure), were statistically significant at a 95% confidence interval. Variable coefficients were listed in Appendix 2. The discretized age variable (Age 60 and above) had the highest coefficient (3.073) indicating old age has the largest influence on CKD and could be a good predictor of CKD too. (Appendix 2)

Based on the ROC curve, the preferred threshold was at ~9.55%. This threshold returned the highest profit and an accuracy of 79%. The model predicted 83 (90.2%) CKD patients correctly while misclassifying 243 persons as CKD patients. It yielded a total profit of $83,600, averaging $69.7 per person.

Figure 1. ROC Curves of Selected Models



## SCREENING TOOL

Our screening tool (Appendix 3) is a short questionnaire crafted to determine whether one is susceptible to CKD. Each respondent will be given a score and if a person scores 8 or above, it suggests that the person has a high likelihood of having CKD and should further visit a doctor or take a confirmatory diagnostic test.

Balancing between the simplicity of the screening tool and the predictive power of the logistic regression model, we shortlisted 8 out of 11 variables from the model to construct the screening questionnaire. Reasons for dropping or changing some variables are as follows:

DBP (Diastolic blood pressure) was not included in the questionnaire as the information can be intuitively deduced from another variable -- Hypertension. A person with hypertension suffers from high blood pressure. Hence, a question about DBP is unnecessary, even though DBP is a statistically significant variable in the logistic regression model.

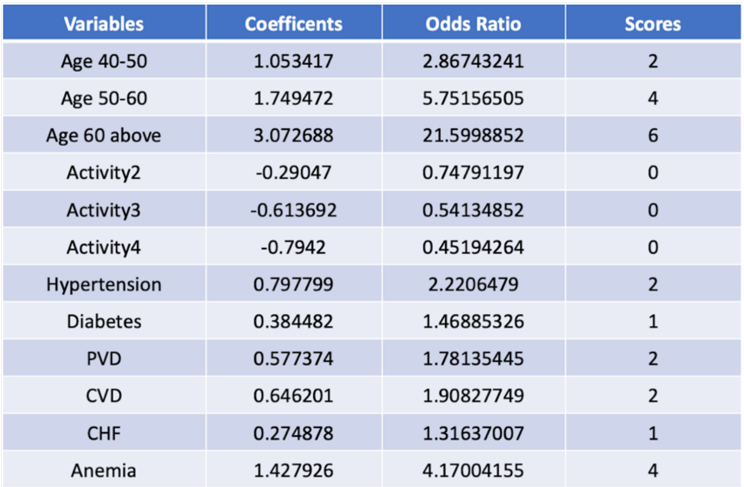
Another statistically significant variable that we left out of the questionnaire is the Hispanic race group. Our models reveal that this race group, either standing alone or together with other race groups, has a negative coefficient. This means that having a Hispanic heritage reduces one's risk of CKD, which contradicts the medical expertise from CDC or National Kidney Foundation. Therefore, to avoid the potential of giving out false information as this data set might be biased, we decided to leave it out of our questionnaire.

With regard to the activities, according to their coefficients and odds ratio, as the level of activity increases, the risk of CKD decreases further. Therefore, we decided to group activity 2/3/4 together as one option of low risk of CKD, while having low activity level (Activity 1) poses a higher risk for CKD. Score-wise, as having activity 2 and above decreases the risk of CKD, we allotted 1 score of having activity 1.

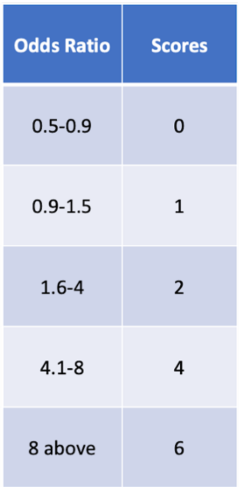
Scores were allotted to the shortlisted variables based on the degree of their individual coefficients. After carefully examining the odds ratio (exponentiated coefficients), they were demarcated into similar scores for those having odds ratio in the comparable range.

In order to determine the questionnaire's cutoff point, or the point at which the questionnaire advises the responder to get tested for CKD, we compare the predictions made by the questionnaire, by the model, and the actual predictions on the data set with CKD results (6000 observations). When the screening tool was applied to that data set, we got a maximum profit of $61.58 per person with an accuracy of 74% at a cut-off score of 6 (7 or above). At a cut-off score of 7 (8 or above), the accuracy increased to 78%. With a tradeoff of $0.6 profit per person, the survey accuracy increases by 4%; though the increase is 1% smaller than the decrease in sensitivity, or true positive rate (Appendix 4). Nevertheless, when tested on the out-sample data set, the difference in true CKD prediction between the regression model and the questionnaire is also at the minimum at cut-off point 7 (Appendix 5). Therefore, a higher cut-off point of 7, though yields a slightly lower profit, ensures the consistency in CKD prediction between the model and the questionnaire.

For reasons given above, a cutoff point of 7 is chosen for the questionnaire. At cutoff point of 7 (8 or above), the ability to correctly predict CKD/non-CKD is consistent across actual vs survey score predictions (Accuracy: 78%, Sensitivity: 81%) and actual vs logistic regression model predictions (Accuracy: 77%, Sensitivity: 85%).

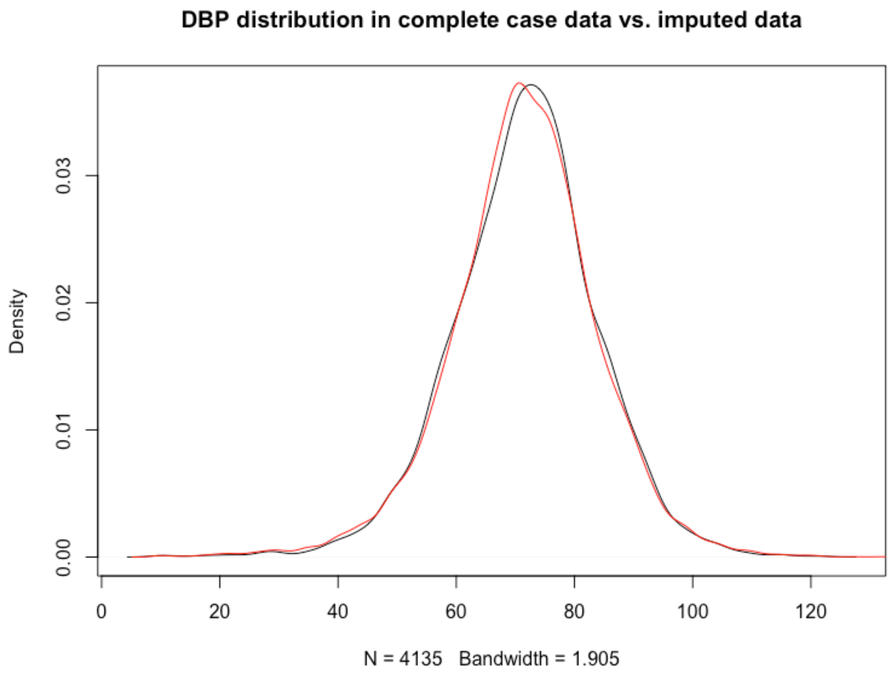


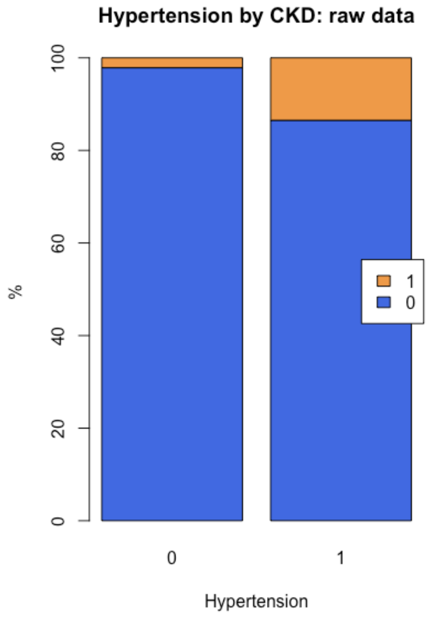
| Variables | Coefficents | Odds Ratio | Scores |
|---|---|---|---|
| Age 40-50 | 1.053417 | 2.86743241 | 2 |
| Age 50-60 | 1.749472 | 5.75156505 | 4 |
| Age 60 above | 3.072688 | 21.5998852 | 6 |
| Activity2 | -0.29047 | 0.74791197 | 0 |
| Activity3 | -0.613692 | 0.54134852 | 0 |
| Activity4 | -0.7942 | 0.45194264 | 0 |
| Hypertension | 0.797799 | 2.2206479 | 2 |
| Diabetes | 0.384482 | 1.46885326 | 1 |
| PVD | 0.577374 | 1.78135445 | 2 |
| CVD | 0.646201 | 1.90827749 | 2 |
| CHF | 0.274878 | 1.31637007 | 1 |
| Anemia | 1.427926 | 4.17004155 | 4 |

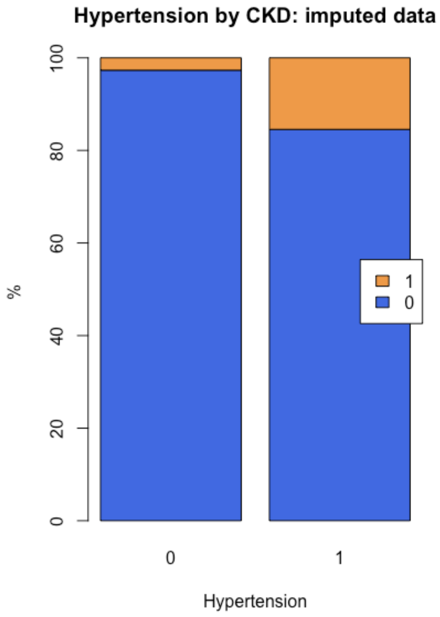


| Odds Ratio | Scores |
|---|---|
| 0.5-0.9 | 0 |
| 0.9-1.5 | 1 |
| 1.6-4 | 2 |
| 4.1-8 | 4 |
| 8 above | 6 |

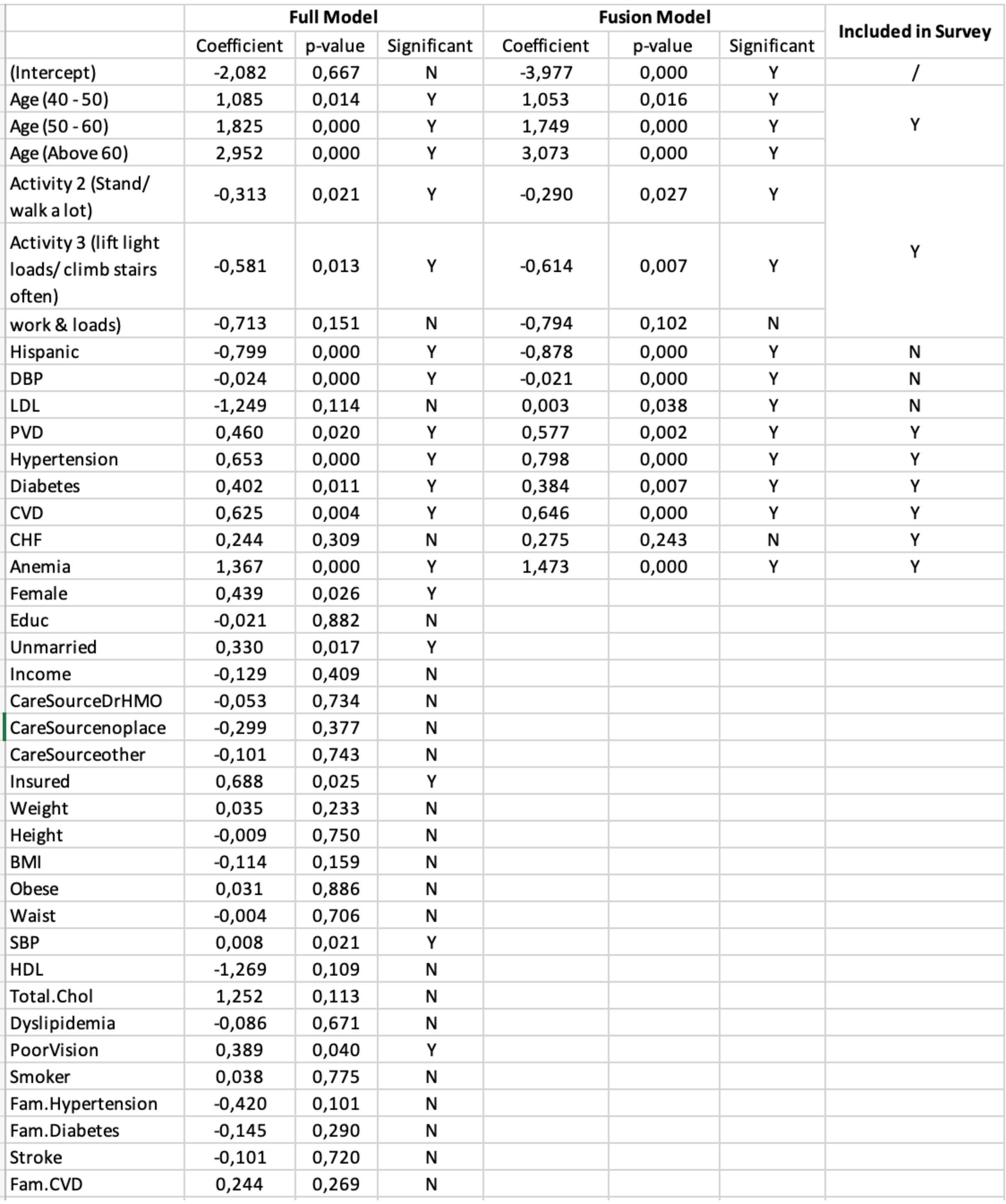
## Appendix 1: Comparison of Variable Distribution before and after Imputation





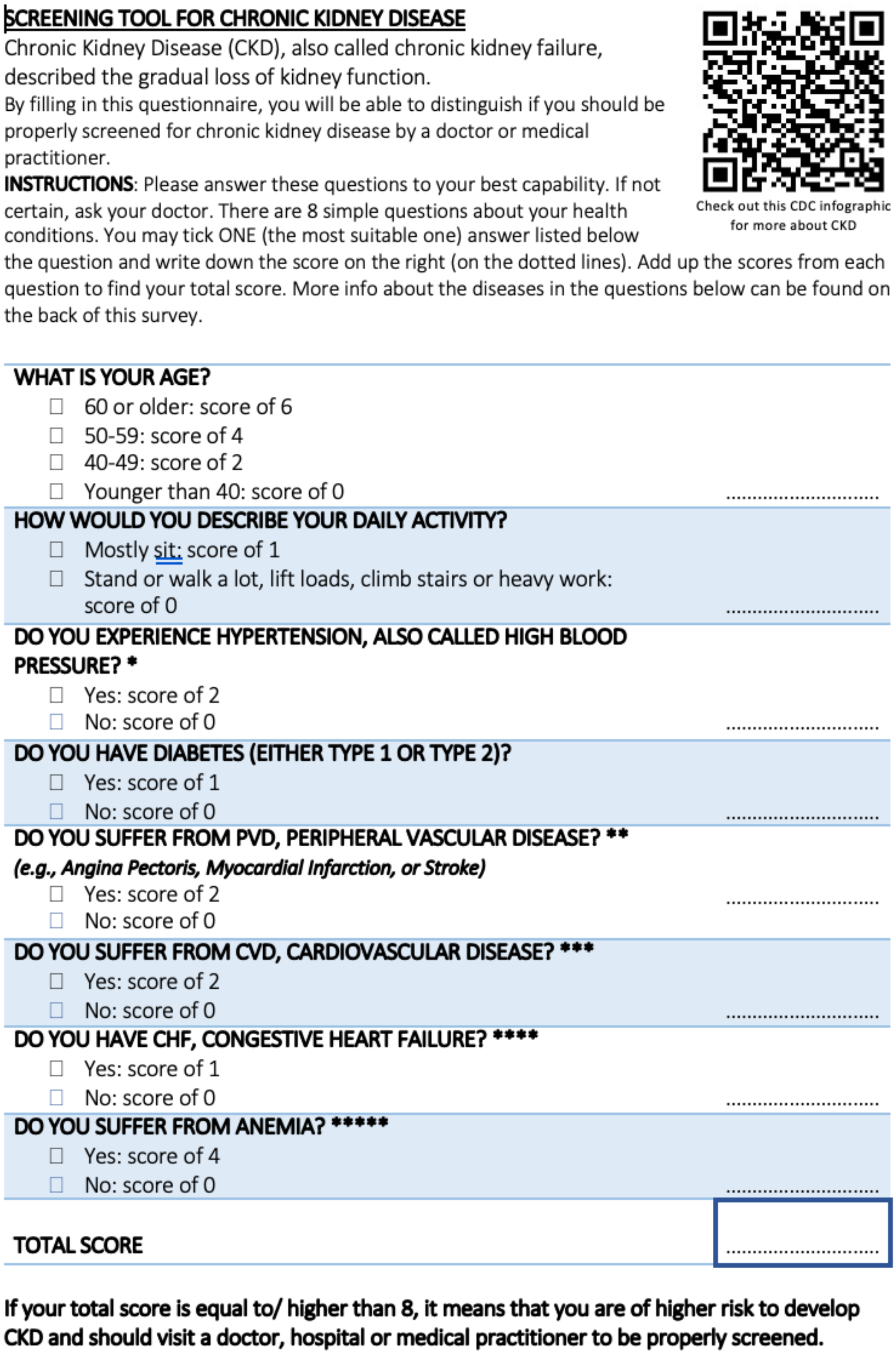


Appendix 2: Coefficients and Statistical Significance of Variables in Full Model and Fusion Model



| | Full Model | | | Fusion Model | | | Included in Survey |
|---|---|---|---|---|---|---|---|
| | Coefficient | p-value | Significant | Coefficient | p-value | Significant | |
| (Intercept) | -2,082 | 0,667 | N | -3,977 | 0,000 | Y | / |
| Age (40 - 50) | 1,085 | 0,014 | Y | 1,053 | 0,016 | Y | Y |
| Age (50 - 60) | 1,825 | 0,000 | Y | 1,749 | 0,000 | Y | |
| Age (Above 60) | 2,952 | 0,000 | Y | 3,073 | 0,000 | Y | |
| Activity 2 (Stand/ walk a lot) | -0,313 | 0,021 | Y | -0,290 | 0,027 | Y | Y |
| Activity 3 (lift light loads/ climb stairs often) | -0,581 | 0,013 | Y | -0,614 | 0,007 | Y | |
| work & loads) | -0,713 | 0,151 | N | -0,794 | 0,102 | N | |
| Hispanic | -0,799 | 0,000 | Y | -0,878 | 0,000 | Y | N |
| DBP | -0,024 | 0,000 | Y | -0,021 | 0,000 | Y | N |
| LDL | -1,249 | 0,114 | N | 0,003 | 0,038 | Y | N |
| PVD | 0,460 | 0,020 | Y | 0,577 | 0,002 | Y | Y |
| Hypertension | 0,653 | 0,000 | Y | 0,798 | 0,000 | Y | Y |
| Diabetes | 0,402 | 0,011 | Y | 0,384 | 0,007 | Y | Y |
| CVD | 0,625 | 0,004 | Y | 0,646 | 0,000 | Y | Y |
| CHF | 0,244 | 0,309 | N | 0,275 | 0,243 | N | Y |
| Anemia | 1,367 | 0,000 | Y | 1,473 | 0,000 | Y | Y |
| Female | 0,439 | 0,026 | Y | | | | |
| Educ | -0,021 | 0,882 | N | | | | |
| Unmarried | 0,330 | 0,017 | Y | | | | |
| Income | -0,129 | 0,409 | N | | | | |
| CareSourceDrHMO | -0,053 | 0,734 | N | | | | |
| CareSourcenoplace | -0,299 | 0,377 | N | | | | |
| CareSourceother | -0,101 | 0,743 | N | | | | |
| Insured | 0,688 | 0,025 | Y | | | | |
| Weight | 0,035 | 0,233 | N | | | | |
| Height | -0,009 | 0,750 | N | | | | |
| BMI | -0,114 | 0,159 | N | | | | |
| Obese | 0,031 | 0,886 | N | | | | |
| Waist | -0,004 | 0,706 | N | | | | |
| SBP | 0,008 | 0,021 | Y | | | | |
| HDL | -1,269 | 0,109 | N | | | | |
| Total.Chol | 1,252 | 0,113 | N | | | | |
| Dyslipidemia | -0,086 | 0,671 | N | | | | |
| PoorVision | 0,389 | 0,040 | Y | | | | |
| Smoker | 0,038 | 0,775 | N | | | | |
| Fam.Hypertension | -0,420 | 0,101 | N | | | | |
| Fam.Diabetes | -0,145 | 0,290 | N | | | | |
| Stroke | -0,101 | 0,720 | N | | | | |
| Fam.CVD | 0,244 | 0,269 | N | | | | |

Appendix 3: Screening Tool



**SCREENING TOOL FOR CHRONIC KIDNEY DISEASE**

Chronic Kidney Disease (CKD), also called chronic kidney failure, described the gradual loss of kidney function.

By filling in this questionnaire, you will be able to distinguish if you should be properly screened for chronic kidney disease by a doctor or medical practitioner.

Check out this CDC infographic
for more about CKD

**INSTRUCTIONS**: Please answer these questions to your best capability. If not certain, ask your doctor. There are 8 simple questions about your health conditions. You may tick ONE (the most suitable one) answer listed below the question and write down the score on the right (on the dotted lines). Add up the scores from each question to find your total score. More info about the diseases in the questions below can be found on the back of this survey.

**WHAT IS YOUR AGE?**

- □ 60 or older: score of 6
- □ 50-59: score of 4
- □ 40-49: score of 2
- □ Younger than 40: score of 0 ............................

**HOW WOULD YOU DESCRIBE YOUR DAILY ACTIVITY?**

- □ Mostly sit: score of 1
- □ Stand or walk a lot, lift loads, climb stairs or heavy work: score of 0 ............................

**DO YOU EXPERIENCE HYPERTENSION, ALSO CALLED HIGH BLOOD PRESSURE? ***

- □ Yes: score of 2
- □ No: score of 0 ............................

**DO YOU HAVE DIABETES (EITHER TYPE 1 OR TYPE 2)?**

- □ Yes: score of 1
- □ No: score of 0 ............................

**DO YOU SUFFER FROM PVD, PERIPHERAL VASCULAR DISEASE? ****

***(e.g., Angina Pectoris, Myocardial Infarction, or Stroke)***

- □ Yes: score of 2 ............................
- □ No: score of 0

**DO YOU SUFFER FROM CVD, CARDIOVASCULAR DISEASE? *****

- □ Yes: score of 2
- □ No: score of 0 ............................

**DO YOU HAVE CHF, CONGESTIVE HEART FAILURE? ******

- □ Yes: score of 1
- □ No: score of 0 ............................

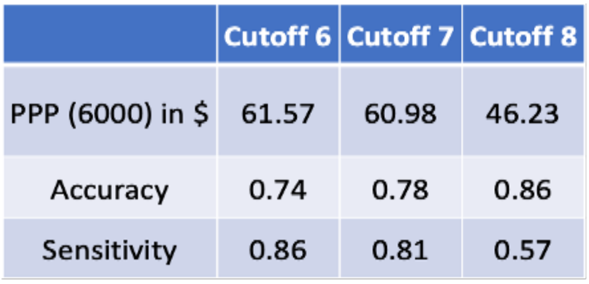
**DO YOU SUFFER FROM ANEMIA? *******

- □ Yes: score of 4
- □ No: score of 0 ............................

**TOTAL SCORE** ............................

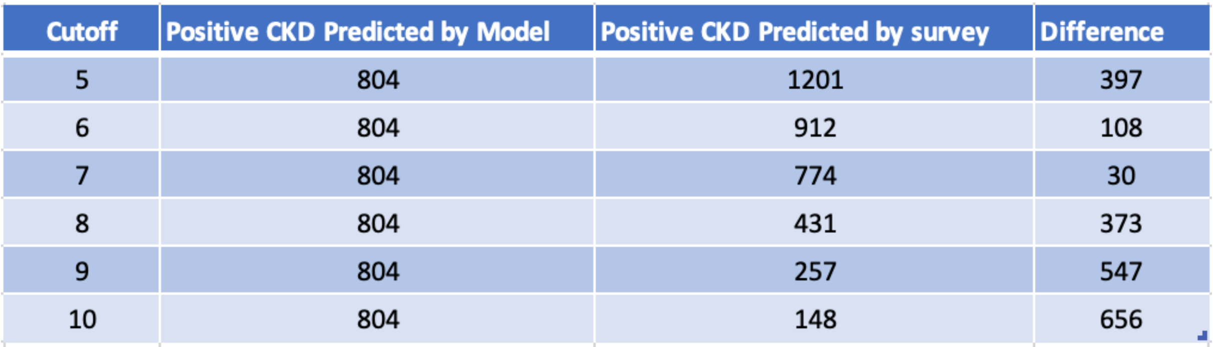
**If your total score is equal to/ higher than 8, it means that you are of higher risk to develop CKD and should visit a doctor, hospital or medical practitioner to be properly screened.**

Appendix 4: Comparison of Actual Profit over different Cutoff Values



| | Cutoff 6 | Cutoff 7 | Cutoff 8 |
|---|---|---|---|
| PPP (6000) in $ | 61.57 | 60.98 | 46.23 |
| Accuracy | 0.74 | 0.78 | 0.86 |
| Sensitivity | 0.86 | 0.81 | 0.57 |

Appendix 5: Comparison of Prediction Power between Screening Tool and Model



| Cutoff | Positive CKD Predicted by Model | Positive CKD Predicted by survey | Difference |
|---|---|---|---|
| 5 | 804 | 1201 | 397 |
| 6 | 804 | 912 | 108 |
| 7 | 804 | 774 | 30 |
| 8 | 804 | 431 | 373 |
| 9 | 804 | 257 | 547 |
| 10 | 804 | 148 | 656 |

[i] Perico, N., & Remuzzi, G. (2012). Chronic kidney disease: a research and public health priority. Nephrology Dialysis Transplantation, 27(suppl_3), iii19-iii26. https://academic.oup.com/ndt/article/27/suppl_3/iii19/1823902?login=true

[ii] Ahmad, F. B., & Anderson, R. N. (2021). The Leading Causes of Death in the US for 2020. JAMA. https://jamanetwork.com/journals/jama/fullarticle/2778234

[iii] https://www.niddk.nih.gov/health-information/kidney-disease/chronic-kidney-disease-ckd

[iv] https://www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/#how_do_i_know_if_i_have_ckd

[v] https://www.cdc.gov/kidneydisease/basics.html

[vi] For a more in detail explaination of MICE, see Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. Int J Methods Psychia Res, 20(1), 40-49. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/

[vii] For a more in detail explanation of missForest method, see Stekhoven, D., J., & Buhlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597

[viii] https://www.kidney.org/atoz/content/minorities-KD

[ix] Pfeifer, P. E., Reynolds, R. S., & Bang H. (2007). Screening for Chronic Kidney Disease(italicized). University of Virginia Darden School Foundation.