Derivation of a Risk Score for prediction of Chronic Kidney Disease (CKD)

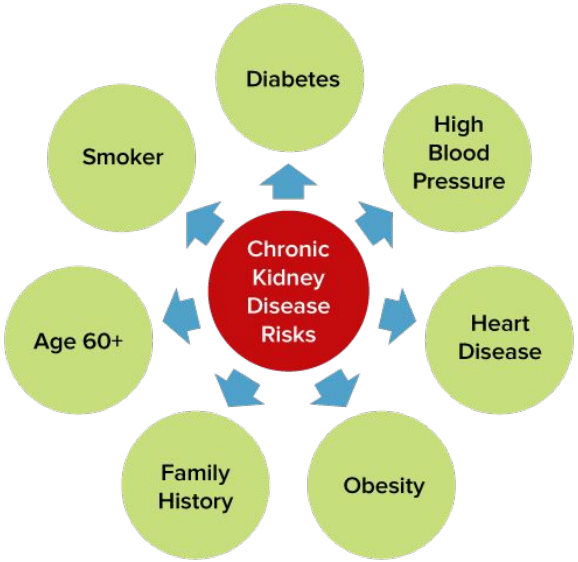
Bryce Parkman & Manan Talwar

CS590W

Motivation

- Chronic kidney disease (CKD) is a leading cause of mortalities, morbidities, and health-care costs in the United States. However, a limited number of recent studies have estimated CKD and its risk factors.

- CKD is considered to be a multifactorial disease related to **sex, age, obesity and smoking, hypertension and cardiovascular diseases** and also to genetic and environmental factors. **Type-1 and Type-2 diabetes**, **hypertension**, the presence of many different **heavy metals either in blood or in urine** and more recently even **demographic factors** such as race are known to be risk factors for acute as well as chronic kidney disease.



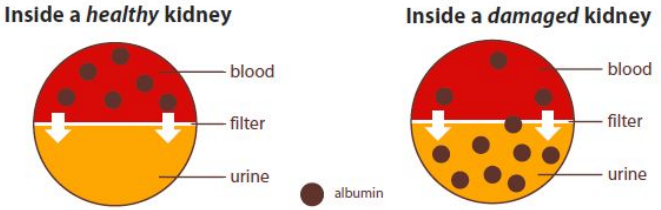
Objectives

The aim of this study is to quantify a patient's kidney disease risk based on a combination of physiological factors including heavy metal presence in blood and urine, urine albumin-to-creatinine ratio (uACR), glomerular filtration rate (GFR), total cholesterol, and blood sugar levels as input variables.

ACR and GFR

- The urine albumin-to-creatinine ratio (uACR) is an important indicator of kidney health. Creatinine is a protein that the liver normally filters into urine, whereas albumin is a protein that the kidney should not filter into urine. A high uACR indicates a potentially damaged liver.

$$\frac{\text{Urine albumin}\left(\text{mg dL}^{-1}\right)}{\text{Urine creatinine}\left(\text{g dL}^{-1}\right)} = \text{UACR}\left(\text{mg g}^{-1}\right)$$

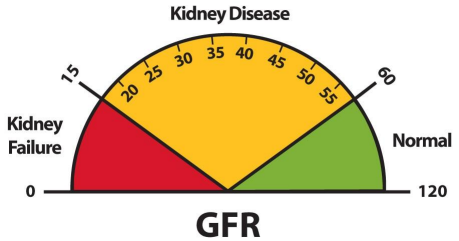


- The glomerular filtration rate (GFR) is also an important indicator of kidney health: it measures the rate at which a patient's kidney is filtering. However this is a complicated and difficult to calculate process, so healthcare professionals have developed a few formulas to estimate this value. We used the CKD-EPI for our research since it generally gives better results over other formulas.

CKD-EPI

$$\text{GFR} = 141 \times \min(S_{Cr}/\kappa, 1)^{\alpha} \times \max(S_{Cr}/\kappa, 1)^{-1.209} \times 0.993^{\text{Age}} \times (1.018 \text{ if female}) \times (1.159 \text{ if African American})$$

*S_{cr} is serum creatinine in mg/dL
κ is 0.7 for females and 0.9 for males
α is -0.329 for females and -0.411 for males
min indicates the minimum of S_{cr}/κ or 1
max indicates the maximum of S_{cr}/κ or 1



ACR and GFR

On their own, ACR and GFR can only give insight into kidney health when they are at extreme values. However, taking them in conjunction can give a much bigger picture. Pictured is a table using both ACR and GFR to estimate kidney health. Here ACR is in mg/mmol, but that easily be converted to our unit of mg/g.

Classification of chronic kidney disease using GFR and ACR categories

GFR and ACR categories and risk of adverse outcomes			ACR categories (mg/mmol), description and range		
			<3 Normal to mildly increased	3–30 Moderately increased	>30 Severely increased
			A1	A2	A3
GFR categories (ml/min/1.73 m^2), description and range	≥90 Normal and high	G1	No CKD in the absence of markers of kidney damage		
	60–89 Mild reduction related to normal range for a young adult	G2			
	45–59 Mild–moderate reduction	G3a[1]			
	30–44 Moderate–severe reduction	G3b			
	15–29 Severe reduction	G4			
	<15 Kidney failure	G5			

Increasing risk

Increasing risk

[1] Consider using eGFRcystatinC for people with CKD G3aA1 (see recommendations 1.1.14 and 1.1.15)

Abbreviations: ACR, albumin:creatinine ratio; CKD, chronic kidney disease; GFR, glomerular filtration rate

Adapted with permission from Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group (2013) KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney International (Suppl. 3): 1–150

Methods: Dataset and Input features

- We used continuous NHANES from 2009 - March 2020 (pre-pandemic).

- ACR can be sourced directly from the dataset, whereas GFR is estimated using the CKD-EPI equation [2] that factors in age, gender, race, and the serum creatinine measurement from the dataset.

- Other input variables included Hb1ac, cholesterol, blood pressure, presence of heavy metals in urine and blood were also directly sourced from the NHANES dataset.

- Additionally, the Questionnaire data includes a Kidney Conditions Questionnaire. We were primarily interested in the variable KIQ022 (“Have you ever been told by a healthcare professional that you had weak or failing kidneys?”)

- Data split is 80/20 for train/test

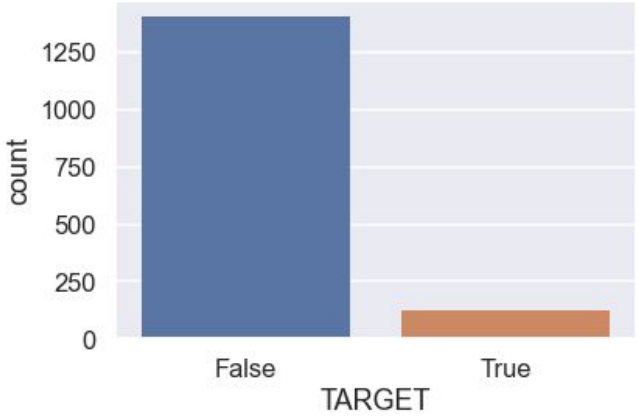
Methods: Data Extraction

- We had to construct our own dataset by merging features across several different NHANES component datasets. These included KIQ022 from the Kidney Questionnaire, Blood Pressure, Metals - Urine, Metals - Blood, Glycohemoglobin, Insulin, and Standard Biochemistry Profile.

- Our dataset was observed to be extremely skewed which needed specific adjustments. Some research suggested that this phenomena is much more common in the practical world!

- Our ‘target’ variable was self-calculated. It is based on the diagnosis criteria for CKD as specified in prior literature and as is used in the medical community these days:

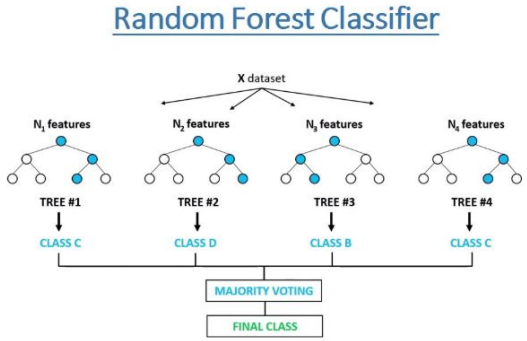
Criteria for CKD: **Either of the following present for > 3 months**	
Markers of kidney damage (one or more)	• Albuminuria (ACR ≥ 30 mg/g) • Urine sediment abnormalities • Electrolyte and other abnormalities due to tubular disorders • Abnormalities detected by histology • Structural abnormalities detected by imaging • History of kidney transplantation
Decreased GFR	GFR <60 ml/min/1.73 m^2



Methods: Design Choices

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

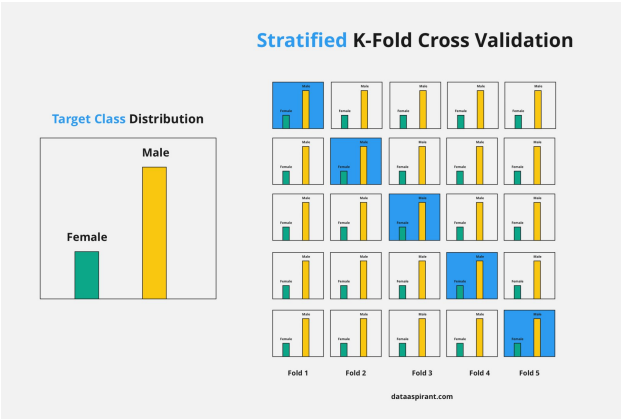
- Skewed dataset ruled out the use of accuracy as a metric. We primarily focused on F1 Scores.
- For developing our risk score and associated classification, we used two models independently: Logistic Regression and Random Forest.
- Skewed dataset necessitated oversampling of the minority (True) class. An implication of the skewed dataset was that the difference between ensemble LR and RF was stark. It was observed that RF dominated by far.
- RF doesn't map 'weights'. Hence, for scores we had considered the probability of a given sample being labelled by the algorithm as True. This provides the 'equation' that predicts the risk score. Other metrics were attempted, but did not yield useful results and were consequently rejected.

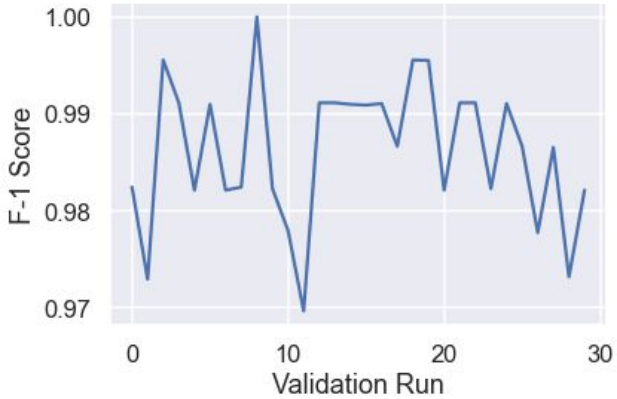


Methods: Design Choices

- We used ensembling within RFs to reduce variance through bagging since individual trees can easily overfit.
- We used Repeated Stratified K-Fold Validation on our training set, selected a model on the basis of our observations and then trained a new model with the training data and evaluated that model on testing data.

- We used three methods for feature selection: Chi2, ANOVA-F and Mutual Information Gain to achieve a comprehensively reduced dimensionality. Our subset included features from all three methods.





Analysis of Observations

- Feature selection resulted in a total of 13 significant features.

  ['LBXSIR', 'LBXTC', 'URXUBA', 'URXUMN', 'LBXGH', 'LBXIN', 'URXUMA', 'URXUCR', 'RIDAGEYR', 'RIDRETH1_1.0', 'LBXSTP', 'LBXSCR', 'RIAGENDR_2.0']

- These correspond to: Iron levels in blood, total cholesterol, Barium levels in urine, Magnesium levels in urine, glycohemoglobin, Insulin, Albumin, creatinine, age, race, total protein, and gender.

- LR ensemble reached a peak F1 score in the range of 75 - 79 percent.

- RF reached a peak F1 score in the range of 89 - 94 percent.

- Owing to the nature of probabilities, risk score is a value between 0 and 1 and was found to be directly proportional to the relative importance of each feature.

Interpretations

- Analysis suggests that Albumin and Creatinine in serum and urine are decisive factors carrying significant importance (expected).

- Glycohemoglobin appears to have a significant contribution.

- Analysis also suggests that presence of metals has some effect. However, this effect is minor and nearly identical across most metals except Manganese which seems to have a more pronounced effect.

References

1] wwn.cdc.gov. 2022. *NHANES Questionnaires, Datasets, and Related Documentation*. [online] Available at: <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?cycle=2017-2020> [Accessed 25 March 2022].

[2] Rate, E. and Health, N., 2022. *Estimating Glomerular Filtration Rate | NIDDK*. [online] National Institute of Diabetes and Digestive and Kidney Diseases. Available at: <https://www.niddk.nih.gov/health-information/professionals/clinical-tools-patient-management/kidney-disease/laboratory-evaluation/glomerular-filtration-rate/estimating> [Accessed 25 March 2022].

[3] Tsai, H., Hung, C., Wang, C., Tu, H., Li, C., Tsai, C., Lin, W., Chen, S. and Kuo, C., 2022. *Associations among Heavy Metals and Proteinuria and Chronic Kidney Disease*.

[4] ucsfhealth.org. 2022. *Lead Levels Blood*. [online] Available at: <https://www.ucsfhealth.org/medical-tests/lead-levels---blood> [Accessed 1 April 2022].