

PRACTICUM'S FINAL PAPER

The Prevalence of Chronic Kidney Disease in Diabetic
Patients with Moderate to High CHA2DS2-VASc Scores
(Male ≥ 2 , Female ≥ 3)

By Kayvan Mivehnejad.

Graduate Student of Biomedical Informatics - Summer 2023,
School of Medicine and Health Sciences,
The George Washington University,

Preceptor Dr. Yin Ying.


Department of Clinical Research and Leadership
The George Washington University



Abstract

Kidney disease often has no symptoms until your kidneys are badly damaged. More than one in seven American adults, or about 37 million, are estimated to have Chronic Kidney Disease (CKD), and 40% of people with severely reduced kidney function (not on dialysis) are not aware of having CKD (CDC, 2023). The underrepresented population of seniors 75 and older with Atrial Fibrillation and Type-II diabetes was selected from the NIH All-of-Us research portal due to the higher risk of vascular outcomes with changes in their renal excretion. American College of Cardiology and American Heart Association (ACC/AHA) scoring system, CHA2DS2-VASc is a potential predictor of stroke; therefore, we investigated the relation of glycemic control with renal outcomes in patients who scored for oral anticoagulation medications by this system. We looked at the correlation between simultaneous reads of the Estimated Glomerular Filtration Rate (eGFR) with Glycated Hemoglobin (HbA1c) and the association of this bivariate with the CKD diagnosis.

The CHA2DS2-VASc ($F \geq 3$, $M \geq 2$) population was split into three groups based on the date and time of CKD diagnosis reflected by the All-of-Us portal's periodical Collection Data Reports (CDR) and measurement dates. A cohort study was designed by the principle of CKD diagnosis and the exposure to eGFR and HbA1c concurrent reads to inspect retrospective measurements recorded before outcome (564 reads in 82 preCKD patients) in comparison with prospective measurements recorded in post-diagnosis dates (1283 reads in 266 NoCKD patients and 987 reads in 144 CKD patients). We denoted a mean eGFR of 58.4 for the entire population with a 46% prevalence of documented CKD outcomes. However, the decreased eGFR (≤ 60), indicating stage-3a mild to moderate kidney damage, was observed in 27% of noCKD cohort in




one test and 13% of noCKD cohort in two tests. Moreover, the indication of stage-2 mild kidney damage ($60 < \text{eGFR} < 90$) was observed once in 95.11% and twice in 67% of noCKD cohort ($\overline{\text{eGFR}}$: ckd = 43.39, noCKD = 69.50, preCKD = 59.46). We observed a relative risk of 4.3 for exposure to decreased eGFR (≤ 60) in CKD cohort.

The mean HbA1c did not show significant statistical differences between cohorts (CKD = 6.72, noCKD = 6.67, preCKD = 6.72), nor showed linear correlation with eGFR in any of the cohorts. Two deep-learning neural networks were built with Python libraries to make binary class CKD predictions after evaluation of the numerical eGFR and HbA1c values in conjunction with nominal indicator coding for race and sex at birth. (Keras AUC= 87%, Accuracy= 82%; PyTorch AUC= 78%, Accuracy= 79%).

Introduction

When kidneys do not work well, toxic waste and extra fluid accumulate in the body, which may lead to heart disease, stroke and death. The CHA2DS2-VASc scoring system is a predefined list of potential predictors of stroke in patients with non-valvular Atrial fibrillation (Afib). The underrepresented population of seniors aged 75 years or older already meet two points in this scoring system despite their gender, so when an additional point due to the onset of type-II diabetes increment to the score, then they qualify for oral anticoagulation medication according to American Heart Association and American College of Cardiology (AHA/ACC). The Cardio-Renal syndrome has been coined to describe a vicious circle in which cardiovascular disease can lead to CKD, which can even independently worsen cardiovascular disease down the line.



Chronic Kidney Disease (CKD) is defined as a Glomerular filtration rate of less than 60 mL/min per 1.73 m² or the presence of kidney damage regardless of cause for more than three months. The gold standard test requires timed urine samples of 24-hrs to show how much fluid is filtered by the kidney using a low-weight polysaccharide inulin. However, urine clearance has many patients' dependent and independent variables, so the nonprofit group of Kidney Disease Improving Global Outcomes (KDIGO) recommends using estimated GFR with either the Modification of Diet in Renal Disease (MDRD) formula or Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation (Edelstein, 2017, p. 3-6). The MDRD formula for eGFR adjusts serum creatinine with body surface area, age, and sex as variables, while the CKD-EPI equation only counts for age and sex.

According to Daugridas (2019), with improved glycemic control, the number of patients going on to more advanced renal outcomes, such as albuminuria and a decrease in GFR are significantly reduced; much of this benefit is related to changes in lifestyle or a smaller number of patients who have developed microalbuminuria to begin with due to fatality. We investigated Glycated Hemoglobin (HbA1C) as a more significant blood examination test than Fasting Plasma Glucose(FPG) for the association of inadequate diabetic control with CKD because it reflects the levels of glucose attached in hemoglobin in the preceding three months with a diagnostic value higher than 6.5% and showing a risk of diabetes (prediabetes) with a value between 5.7% and 6.4%. Currently, the recommended target for diabetic patients with CKD is HbA1c of < 7% (53 mmol/mol) (Gunton's 2014 study as cited in Daugridas 2019).



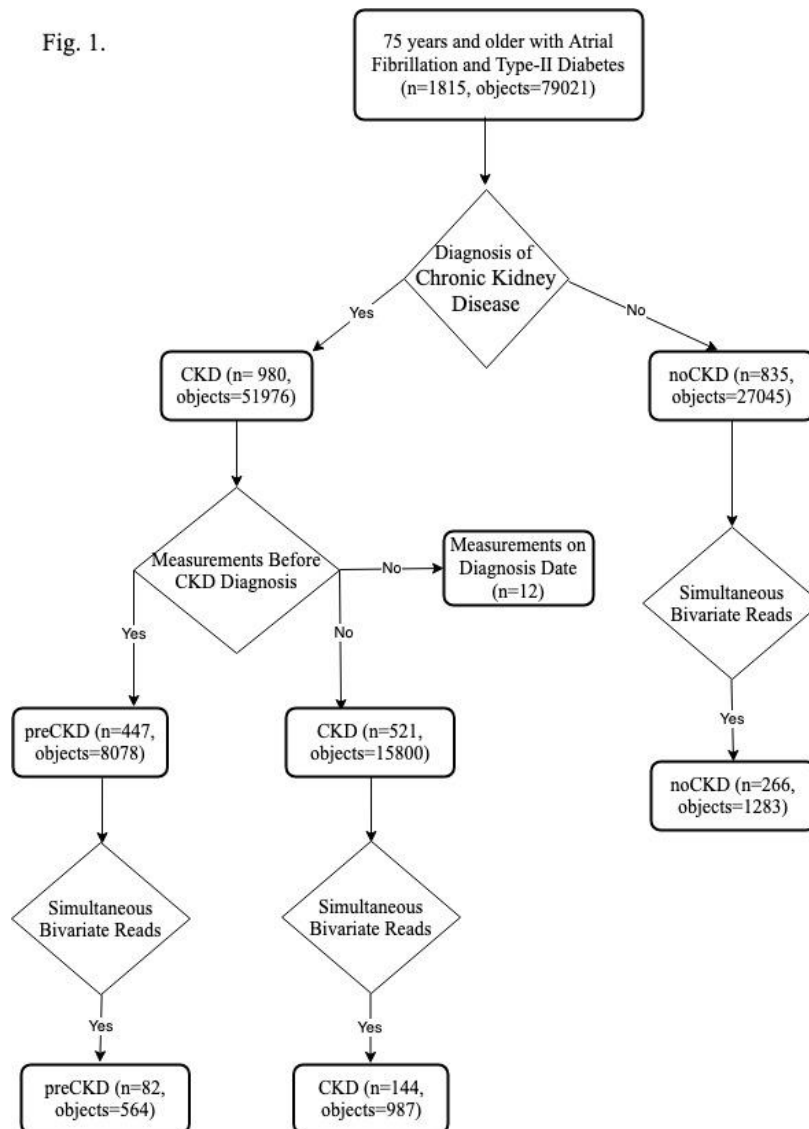
Methods


The NIH All-of-Us Research platform was engaged for evaluating real patients' health records, and we selected the underrepresented population of adults aged 75 and older with type-II diabetes and Afib to meet the CHA2DS2-VASc score of two or higher for men; three or higher for women. The R programming language was used to combine persons' demographic database with conditions and measurements to build the initial dataset of 79021 objects (test results) for 1815 patients. This dataset was divided into two subsets based on the outcome of CKD diagnosis in electronic health records, then we wrangled these two subsets for more than one measurement per person to build a bivariate simultaneous dataset of eGFR and HbA1c reads. This data

munging reduced the total number of qualified objects to 2834 concurrent reads for 492 patients (CKD = 144, preCKD = 82, noCKD = 266) (Fig. 1.).

The one-way ANOVA test was used to analyze the difference in mean of eGFR and HbA1c for total of 2834 laboratory test results based on independent variable (factor) of CKD status with three levels. Consequently, we used two deep learning neural networks from PyTorch and Keras(Tensorflow) libraries distributed in Python programming language for the purpose of CKD status prediction. We instantiated the Multilayer Perceptron (MLP) model class for binary


Fig. 1.





classification of CKD status based on eGFR and HbA1c as two numerical variables, but we also used indicator coding for inclusion of sex at birth (Female=1, Male=0) and developed a coding system for race with a value of one for either Black, Asian or others (More than one race or unanswered) as their respective features in contrast to saving the value of zero in these three features for showing the white race (Fig. 2.).

The combination of CKD and noCKD datasets with 2270 objects was split into 1521 training records (67%) for the network to learn weights and 749 test objects (33%) for model evaluation. A fully meshed network of nodes in three layers was designed with rectified linear unit (Relu) activation function in the first two layers with “Kaiming-He” weight initialization and sigmoid activation function in the output layer to generate probability predictions. The training dataset was built from epochs of shuffled 32x6 variable batches and fed into the network by clearing the gradients of the previous training epoch for each current of Stochastic Gradient Descent (SGD) optimization, then forward passing the new epoch into the model to calculate loss using Binary Cross Entropy Loss (BCELoss) in comparison to actual values. The backpropagation of the calculated loss through the model for coefficient(weight) modifications reduced the loss with step sizes of 0.01 and momentum of 0.9 within SGD configurations.



	egfr	hba1c	sab	race_black	race_asian	race_other	ckd_status
0	58.0	6.8	0	0	0	0	0
1	60.0	6.4	0	0	0	0	0
2	60.0	6.2	0	0	0	0	0
3	60.0	5.8	0	0	0	0	0
4	77.0	5.8	0	1	0	0	0

Fig. 2.



	egfr	hba1c	sab	race_black	race_asian	race_other	ckd_status
2265	66.0	5.8	1	0	0	0	1
2266	64.0	6.3	1	0	0	0	1
2267	63.0	6.8	1	0	0	0	1
2268	60.0	6.0	1	0	0	0	1
2269	57.0	6.1	1	0	0	0	1

The tabular dataset of CKD merged with noCKD features was saved in All-of-Us secure workspace bucket, and it was redrawn for further neural network development in the Keras library. Similarly, we cleaved the dataset for training the model and used sigmoid and Relu activation functions in four dense layers of nodes with Adam optimization algorithm.

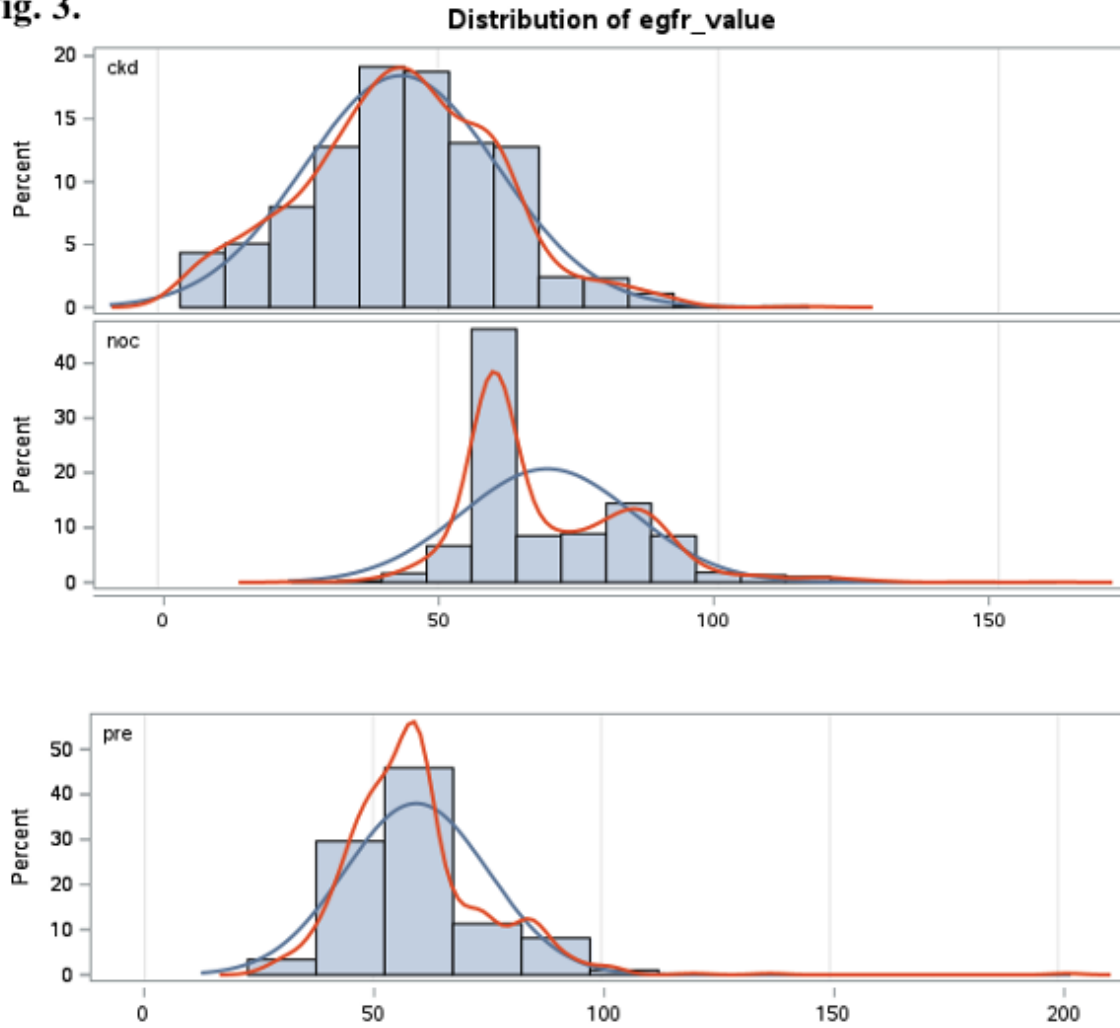
Results

At the day that we collected data from underrepresented population of 75 and older from All-of-Us web portal, the average age of 266 persons in the CKD cohort was 82 years, while similarly preCKD cohort showed an average of 81 years, and noCKD cohort returned 80 years. We noticed that proportion of Black/African in CKD & preCKD cohorts (12.5% - 19.51%) was considerably higher in comparison with noCKD cohort (6.76%). The distribution of race, sex at birth and ethnicity is shown in Table. 1.

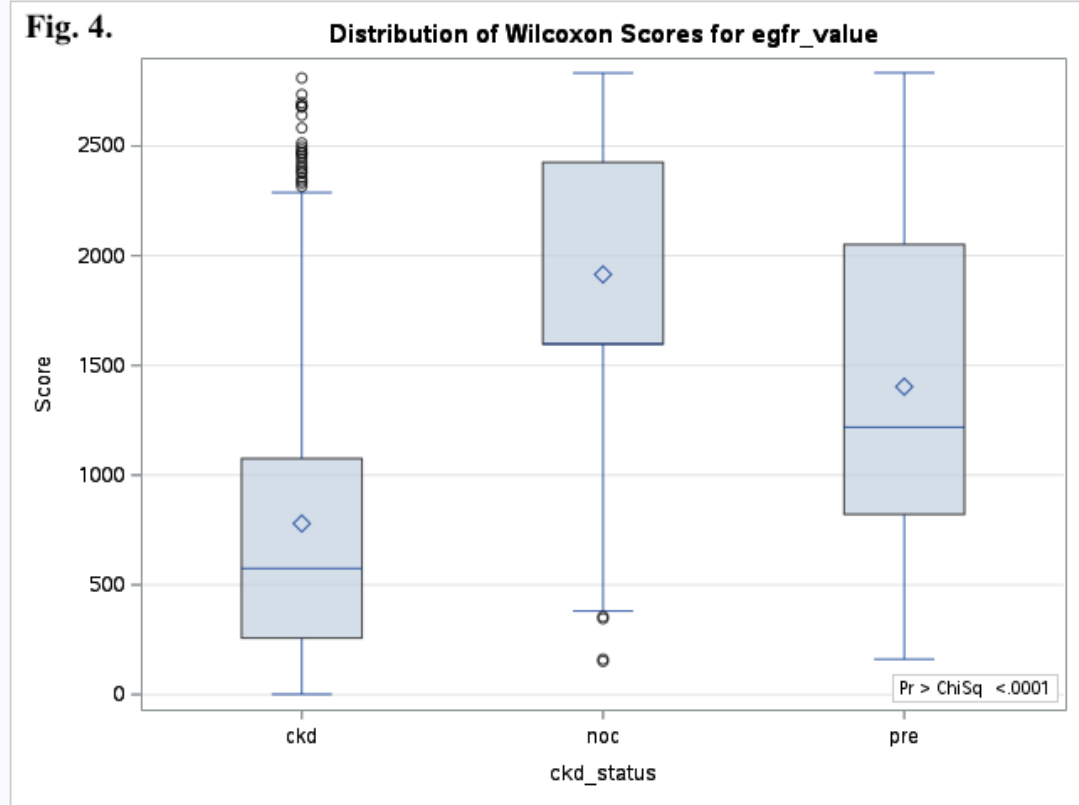
Table. 1.	NOCKD (n=266)	PRECKD (n=82)	CKD (n=144)
Sex At Birth			
Male	140 (52.63%)	41 (50%)	73 (50.69%)
faemale	122 (45.86%)	39 (47.57%)	67 (46.53%)
Skipped	4 (1.5%)	2 (2.44%)	4 (2.78%)
Race			
White	184 (69.17%)	52 (63.41%)	91 (63.19%)
Black	18 (6.76%)	16 (19.51%)	18 (12.5%)
Asian	5 (1.87%)	1 (1.22%)	1 (0.69%)
More than one	2 (0.75%)	1 (1.22%)	2 (1.39%)
Ethnicity			
No Answer	57 (21.43%)	12 (14.63%)	32 (22.22%)
Hispanic	51 (19.17%)	6 (7.32%)	27 (18.75%)
Not Hispanic	204 (76.69%)	70 (85.37%)	111 (77.08%)
Skipped	11 (4.14%)	6 (7.32%)	6 (4.17%)

The one-way ANOVA test assumption for homogeneity of eGFR variance between three cohorts was rejected with Levene's test (F-Value = 3.34, P-value = 0.0354), and correspondingly observing the frequencies of test results in the histogram of noCKD cohort revealed distinct bimodal peaks for 46.22% of eGFR values at 60 and 14.42% of eGFR values at 84, so we used the non-parametric ANOVA (Kruskal-Wallis test) to investigate the differences in sorted ranks of eGFR medians ($p < 0.0001$).

Fig. 3.



The mean score of eGFR was significantly different between three cohorts. The hinge plot in Fig.4. shows each cohort's median, interquartile range, minimum and maximum scores. Even though there are some overlaps throughout the range of values in the three cohorts, but it is clear that people with diagnosed CKD had lower median. We also observed the mean eGFR of 43.39 mL/min/1.73 m² with std-dev of 17.32 in 987 CKD patients, 69.5 mL/min/1.73 m² with std-dev of 15.41 in 1283 noCKD patients and 59.46 mL/min/1.73 m² with std-dev of 15.73 for 564 preCKD patients.

Fig. 4.

Similarly, we used the one-way ANOVA test to check the means of HbA1c results between three cohorts. The Levene's test for homogeneity of HbA1c variances across three cohorts was passed (F-Value = 0.06, P-value = 0.9372), and the mean square among groups was 0.8 with two degree of freedom in the numerator, and mean square within groups was 1.33 for 2831 degree of freedom in the denominator. The total F-statistics was 0.6 (p-value 0.5482) which confirmed the null hypothesis for equity of HbA1c means across three cohorts (Table. 2A.). The hinge plot in Fig. 5. shows adjacency of medians across cohorts with right-skewed rank outliers in all cohorts.

Table. 2.
(A)

Dependent Variable: hba1c_value

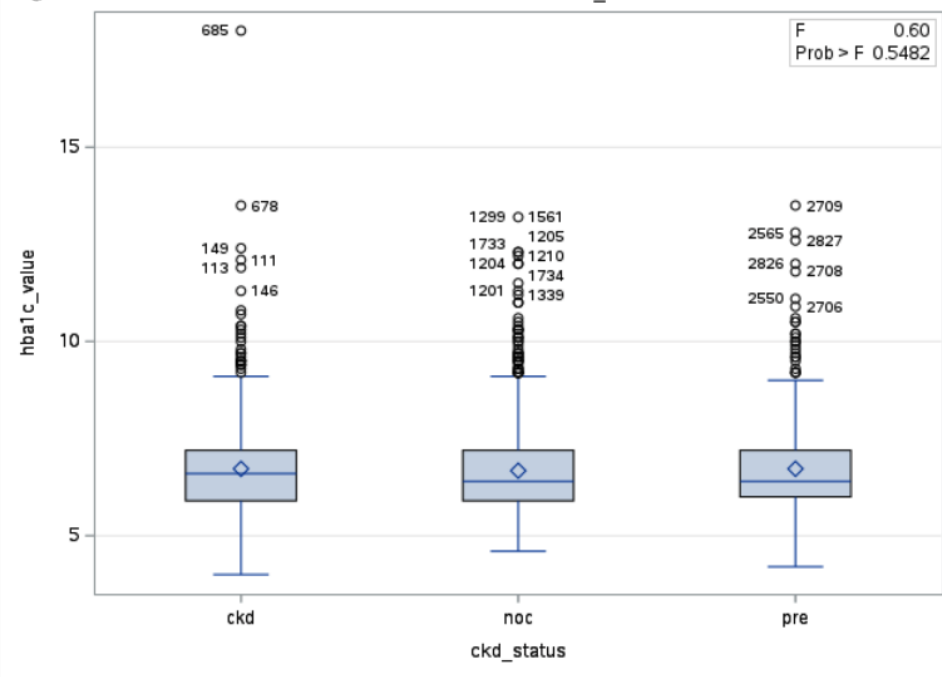
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.601801	0.800900	0.60	0.5482
Error	2831	3771.050484	1.332056		
Corrected Total	2833	3772.652284			

(B)

Level of ckd_status	N	hba1c_value	
		Mean	Std Dev
ckd	987	6.72143870	1.14294012
noc	1283	6.67393609	1.15313986
pre	564	6.72214539	1.17576996


Fig. 5.

Distribution of hba1c_value



Discussion

The Renin-Angiotensin-Aldosterone system (RAAS) plays an important role in maintaining blood pressure. Macula Densa cells and the sympathetic nervous system trigger the Renin hormone; consequently, the Angiotensin II and increased concentration of solute in plasma triggers the hypothalamus to transport Antidiuretic Hormone (ADH), to pituitary master gland so



it could be released to blood stream causing thirst and reabsorption of water in renal distal tubules. The long-term control of blood pressure requires the maintenance of extracellular fluid (ECF) by the renal system and antidiuretic fluid regulation to retaining water in the bloodstream that will increase the blood pressure. The inclusion of hypertension as one of the components of the CHA2DS2-VASc scoring system could be surrogated by the ADH measurements recorded on the same day and time as other biomarkers in this cohort study. However, we only found one patient in our cohorts that could extend current bivariate simultaneous reads with ADH measurement, so we decided to exclude this variable from our study of CKD risk. Similarly, the query for the renal failure predictive ratio of Blood Urea Nitrogen (BUN) over plasma creatinine only returned four simultaneous reads alongside eGFR and HbA1c results. The ratio of microalbumin(mg/dl) over creatinine(g/dl) in urine returned five persons from the All-of-US records, so we did not use these physiologically plausible test results in our study.

The development of multivariate logistic regression model was influenced by limitations described above, so we extended matrices of numerical eGFR and HbA1c biomarkers with nominal indicator codes of race and sex at birth to train models that could eventually predict the status of CKD. The evaluation of model fitted with the merged dataset of noCKD with CKD cohorts in PyTorch and Keras (tensorflow) provided the performance metrics described in Fig. 6. confusion matrix & Table 3 ratios against actual diagnosis results.

Fig. 6.

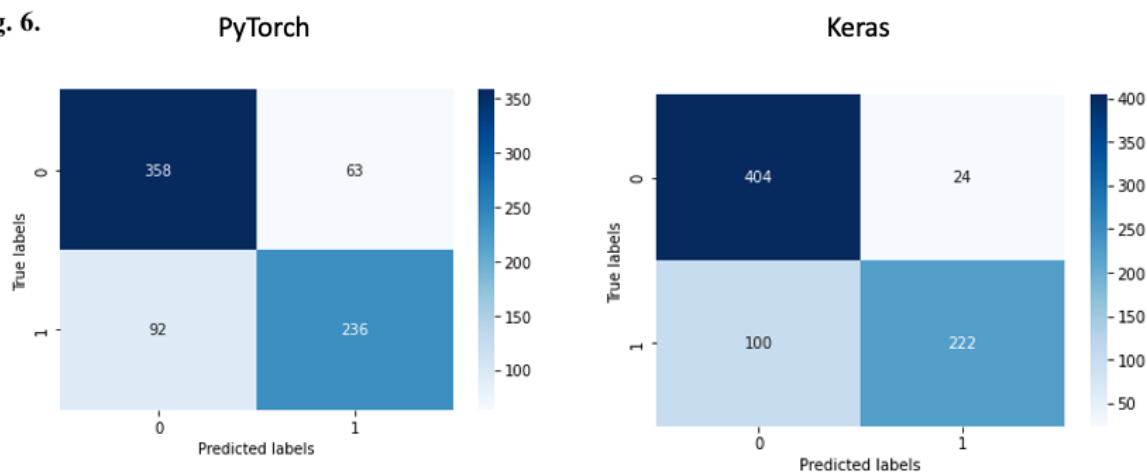


Table. 3.	Sensitivity	Specificity	Accuracy	PPV	NPV	AUC
PyTorch	0.72	0.85	0.79	0.79	0.8	0.78
Tensorflow keras	0.69	0.94	0.82	0.9	0.8	0.87

The deployment of the linear model function in R programming language showed that, there was a no correlation between eGFR and HbA1c values in neither of the cohorts. In CKD cohort, when HbA1c was known, the value of eGFR changed only 0.086% (Rho of CKD = -0.02938, R-Squared = 0.0008637), and for preCKD cohort when HbA1c was known the value of eGFR changed 0.2% (Rho of preCKD = 0.0455, R-squared = 0.002072) and finally for noCKD cohort when the value of HbA1c was known we noticed 0.0001722% change in eGFR (Rho of noCKD = -0.001312, R-squared= 1.722e-06)(Fig. 7.).

Fig. 7A.

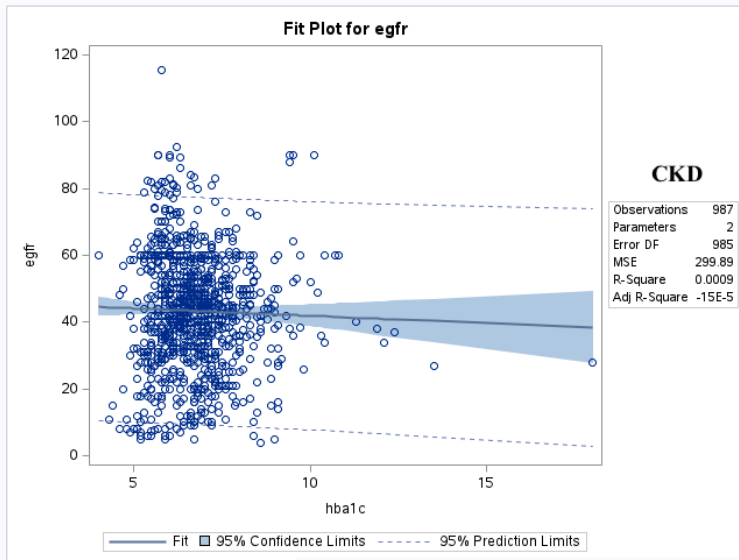


Fig. 7B.

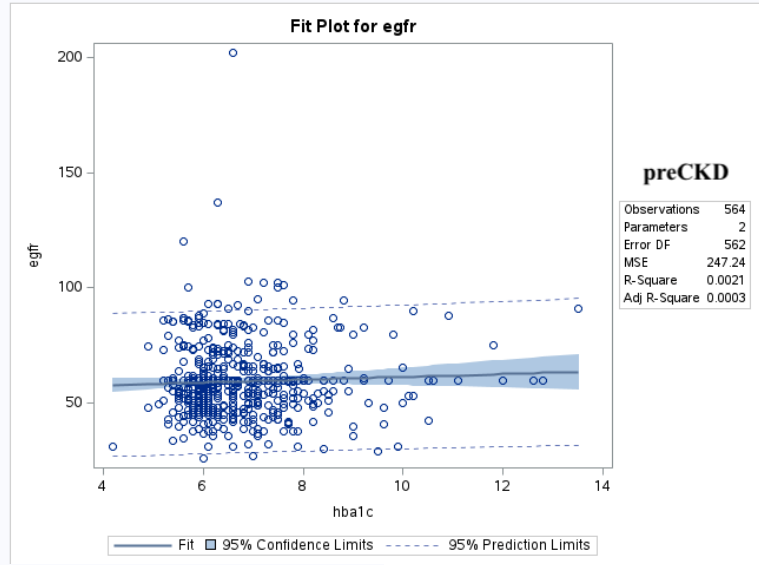
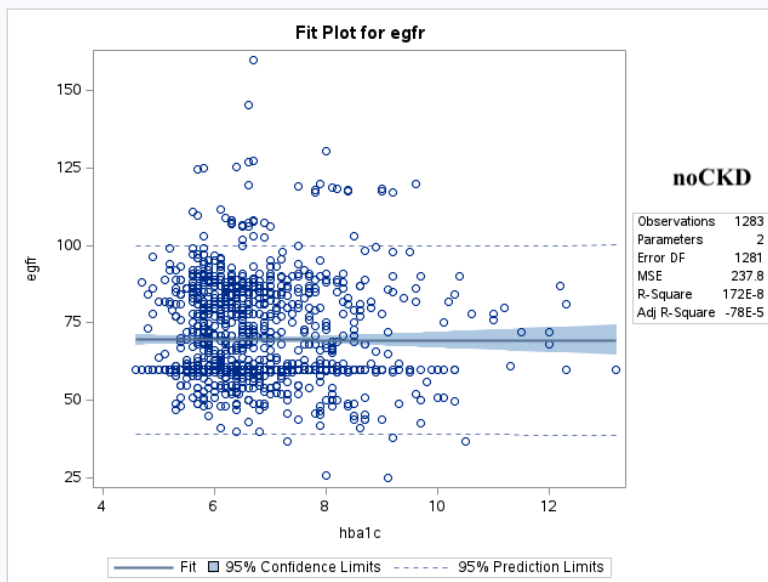


Fig. 7C.






Conclusion

The HbA1c as a long-term surrogate of the hyperglycemia management showed equality of means across the three cohorts; However, the bimodal peaks in the frequency of eGFR value was an alert to extract decreased $eGFR \leq 60$ as a key marker for stage-3a mild to moderate kidney damage in 27% of noCKD cohort once and 13% of noCKD patients with two tests. Respectively, the decreased $60 < eGFR < 90$ as a key marker for stage-2 mild kidney damage returned one test in 95.11% of these undiagnosed patients and two tests in 67% of noCKD cohort. Therefore, CHAD2S2-VASc score ($F \geq 3$, $M \geq 2$) could be a supportive score system for the opportunistic screening of asymptomatic patients and monitoring of renal impairment (Mean of $eGFR = 58.40$) at early treatable stage. When subclinical nature and non-apparent symptoms of chronic kidney disease remain untreated, it could consequently, increase risk of vascular adverse events. We observed relative risk of 4.3 for exposure to decreased $eGFR (\leq 60)$ in CKD cohort.

More than fifty substances that affect blood coagulation have been found in human blood and tissues. Some promote coagulation, called procoagulants, and others inhibit coagulation, called anticoagulants. In the bloodstream, anticoagulants normally predominate, but when the vessel is ruptured or blood is immobile for a long time, then procoagulants become activated and override. Vitamin K is required for the activation of coagulation proteins produced by the liver, such as fibrin (factor-I), thrombin (factor-II) and factor-X (Stuart-Prower)(Hall & Hall, 2020). CKD, as one of the leading causes of death that alter blood clotting, needs to be included in the decision for anticoagulation medication. If we continue monitoring these cohorts in a case-control study, then we can investigate the prevalence of thromboembolic complications with the inclusion of other factors such as anticoagulation agents. Currently, 67.91% of CKD cohort and 64.59% of



noCKD cohort are prescribed with oral anticoagulant agents either in the NOAC form (apixaban, rivaroxaban & dabigatran) or VKA(warfarin).



References

Centers for Disease Control and Prevention(CDC), 2023, Chronic Kidney Disease Basics,
<https://www.cdc.gov/kidneydisease/basics.html>

Daugirdas, J. T. (2019). Handbook of Chronic Kidney Disease Management. (Second edition.).
Wolters Kluwer Publishing.

Edelstein, C. L. (2017). Biomarkers of Kidney Disease (2nd. edition). Elsevier Publishing.

Hall, J.E. & Hall, M.E. (2020). Guyton and Hall Textbook of Medical Physiology E-Book, 14th Edition,
Elsevier. <https://bookshelf.health.elsevier.com/books/9780323640060>

Tam, A. (2023). Deep Learning with PyTorch. Machine Learning Mastery.