Analyzing the Influence of Virtual Care Management on Patient Health Outcomes and Risk Stratifying Chronic Kidney Disease Patients for Stage Degeneration

Yonatan Ashenafi [1], Debabrata Auddya [2], Sayoni Chakraborty [3], Zachary Dana [4], Manraj Singh Ghumman [5], Dimitri Lopez [6], Huy Pham [7], Kriti Sehgal [8], Frederick Senya [9], Abiodun Sumonu [10], Lan Trinh [11], Yuehui Xu [12], Li Zhang [13]

[1] *Worcester Polytechnic Institute*
[2] *University of Delaware*
[3] *The University of Texas at Dallas*
[4] *Vironix Health*
[5] *University of Pittsburgh*
[6] *Rensselaer Polytechnic Intitute*
[7] *Mississippi State University*
[8] *The Ohio State University*
[9] *University of Vermont*
[10] *The University of Alabama*
[11] *Tulane University*
[12] *Indiana University-Purdue University*
[13] *The Citadel*

(*Communicated to* MIIR *on 23 August 2024*)

Study Group: Mathematical Problems in Industry Workshop, University of Vermont, June 25–29, 2024

Communicated by: Taras I. Lakoba, University of Vermont

Industrial Partner: Vironix Health

Presenter: Dr. Sumanth Swaminathan

Industrial Sector: Biomedical/Healthcare; Data Analysis

Tools: Python

Key Words: data cleaning, data visualization, multiple linear regression, k-means clustering, Gaussian process regression, decay models, principal component analysis

1 Introduction

The primary function of the kidneys is to filter and remove waste from the body thereby maintaining several features of physiological homeostasis. Kidney health can be estimated by the Glomerular Filtration Rate (GFR) which represents the total of the filtration rates of nephrons in the kidney. It is typically measured in ml/min. A lower GMR is associated with lower kidney function while the normal values remain in the range of GFR > 90 ml/min. Chronic Kidney Disease (CKD) is a pathological condition where the kidneys become degraded over a period of time. CKD is considered a public health problem characterized by a silent disease progression and gradual decline which can lead to End Stage Kidney Disease (ESKD) reflecting kidney failure (GFR < 15 ml/min) and requiring regular dialysis or a transplant. According to the CDC in 2021, 15% of adults in the US (more than 1 in 7 people) have CKD and around 786,000 people are living with ESKD [1]. Early intervention in CKD patients can lead to an improved quality of life by slowing down the progression of CKD and reducing healthcare expenses [3]. According to the United States Renal Data System 2022 Annual Data Report Medicare spending for beneficiaries with CKD (not including ESKD) ages 66 or older exceeded $75 billion in 2020, representing 25.2% of Medicare spending in this age group and Medicare-related spending for beneficiaries with ESKD totaled $50.8 billion in 2020 [2]. Mathematical modelling has been of critical importance in developing inexpensive methods for estimating kideny health. One important contribution has been in evaluating Estimated Glomerular Filtration Rate (eGFR) rather than directly estimating GFR using more expensive techniques like exogenous filtration markers [9, 6]. Predictive models and dynamic risk stratification algorithms can aid in identifying patients at high risk of CKD degeneration (e.g. through changes in controller/BP medications and adjustments in diet, sleep, and exercise) as well as insights on how to manage and screen outpatients with CKD [11]. The Kidney Failure Risk Equations are a widely accepted tool for predicting the probability of kidney failure in patients with stage G3-G5 CKD. These equations use either 4 variables or the more accurate version with 8 variables at a point in time to evaluate the probability of kidney failure in 2 to 5 years [10].

2 Literature Review

In order to investigate the challenges associated with accurate prediction of the risk of kidney diseases a number of machine learning models have been implemented with various datasets which identify parameters relevant to CKD. In Bai et al.'s [3] study, five machine learning models were tested to predict end-stage kidney disease (ESKD) in CKD patients over five years. The random forest model performed best overall, with an area under curve (AUC) of 0.81, while ML models showed higher sensitivity than the traditional Kidney Failure Risk Equation (KFRE). The results indicated that ML models can be more effective for early screening of patients at risk of ESKD.

Xiao et al. [11] supported Bai et al. [3] by showing that linear and ensemble-based learning models performed best in predicting 24-hour urinary protein outcomes for CKD patients. Using nine predictive models and various blood tests and demographic features, they found that logistic regression was the top performer with an area under receiving

operator characteristics curve (AU-ROC) of 0.873. Key predictors of CKD progression included albumin, serum creatinine, triglycerides, low density lipoprotein (LDL), and estimated glomerular filtration rate. Elastic Net had the highest sensitivity, while XGBoost had the highest specificity.

To model temporal distributions classical statistical methods have been used by Ye et al. [12] who utilized Cox proportional hazards regression to develop a predictive model for three-year adverse outcomes in East Asian CKD patients. The model, visualized as a nomogram, demonstrated high discrimination with C-statistics of 0.90, 0.91, and 0.83 for development, internal validation, and external validation datasets, respectively. Calibration plots indicated strong agreement between predicted and observed outcomes. Decision curve analysis highlighted the nomogram's superior clinical value over eGFR alone, particularly at critical threshold probabilities. Understanding the parameters relevant to modeling and tracking risk associated with CKD involves large volume of datasets. While the current modeling approaches use a well established dataset, we have been able to identify several other datasets with well defined parameters which promises better prediction capabilities.

The KFRE is a tool that helps predict the risk of end-stage renal disease (ESRD) and the need for dialysis or a kidney transplant within 2–5 years for people with CKD. It uses four key factors: age, sex, urine albumin-to-creatinine ratio (ACR), and eGFR. Recently, a study by Major et al. [7] aimed to validate how well these predictors work. They found that the recalibrated KFRE accurately predicted the risk of ESRD at both 2 and 5 years when used in primary care settings. Based on these findings, the authors suggested that using this model in primary care could help reduce unnecessary referrals to specialists and ensure that patients who are at higher risk of developing ESRD are referred earlier. This could improve patient outcomes by providing timely interventions. Additionally this data contained parameters such as Albumin-to-creatinine ratio (ACR) and EPI-EGFR which are critical to CKD detection.

We have additionally identified another robust dataset from the works of Mota-Zamorano et al. [8] which highlight the role of arachidonic acid (AA)-derived eicosanoids on Diabetic Kidney Disease (DKD). Their results demonstrate that levels of vasoactive eicosanoids in plasma and urine are linked to kidney function, as shown by protein levels in the urine (proteinuria) and estimated glomerular filtration rate (eGFR). More importantly, we found significant differences in these levels between patients with diabetic kidney disease (DKD) and non-diabetic individuals. These findings support that AA-derived metabolites in plasma and/or urine could be helpful in diagnosing DKD, a condition that still lacks reliable biomarkers.

Due to the multifactorial and complex nature of CKD, recent advances in predicting the factors relevant to predictive diagnosis uses supervised learning. In a recent study, Dritsas and Trigka [4] analyzed the application of machine learning methods to forecast the likelihood of chronic renal disease. They highlighted that machine learning could enhance the accuracy of CKD risk prediction models due to its ability to handle large datasets and identify complex patterns. Additionally, the authors explored methods for determining the appropriate diet plan for CKD patients by using various classification techniques such as multi-class decision trees, multi-class decision forests, multi-class logistic regression, and multi-class artificial neural networks (ANN). Their findings revealed

that the multi-class decision forest achieved the highest accuracy (99.17%) compared to the other models. Their study also addressed some drawbacks of predictive methods, including the necessity for high-quality data and the risk of overfitting.

3 Virtual Care Management's influence on Patient Outcomes

In order to analysis and give more insights into the performance of virtual care management by Vironix, we use four datasets collected by Vironix. These datasets are patient's Demographics, Observation records, Encounter records, and Questionnaire results. The Demographics dataset contains the patient's profile data such as age, gender, height, weight and disease history. The Observation datasets contains all the entries recorded by the patient on the Vironix platform. These entries include body mass, heart rate, blood pressure, oxygen saturation and other health indexes that could be recorded without a medical provider. The Encounter dataset contains the records of every doctor visit of each patient, including the start date, end date, and encounter type for each visit. The Questionnaire dataset contains information that is filled by patients when they have any symptoms. We used these datasets to analysis the compliance of patients and the influence of Vironix's virtual care management on patients.

In the first dataset, we were provided with the demographics and medical history of the patients enrolled with virtual care management by Vironix. First, we parsed the dataset originally in json format to obtain all the relevant information in the columns of our dataframe. We cleaned the original dataset and created sub-datasets only containing demographics information with the goal of understanding Vironix's customer demographics. In the following sections, we will show the results we got from combining it with the Observation dataset and the Encounter dataset.

3.1 Observations

The original observation data set has all type of health information mixed together, which means values of different meaning and units are combined together. Therefore, we separated the observation dataset into multiple small dataset based on the type of each entry, such as body mass and blood pressure. After dividing the observation data into small datasets, we visualized each small dataset to have an overview of the observation data (Figure 1). From Figure 1, we noticed that most of the observations were obtained by people who has regular health index, which is expected since most people are in average healthy condition. From the histogram of body mass, we can also notice that there are more data points on $\geq$ 150kg, which means overweight patients are more likely to record their weight.

On the other hand, we also want to see how population distribution is along different number of observations. Firstly, we tried to plot the relationship between the average health index and the number of observation entries we have (Figure 2).

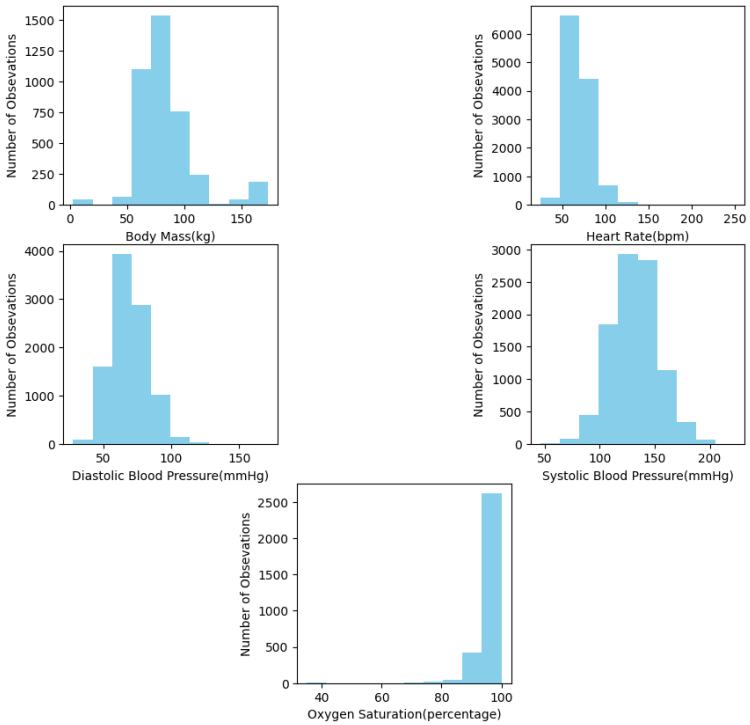


Figure 1. Histograms of number of observations over different health indexes

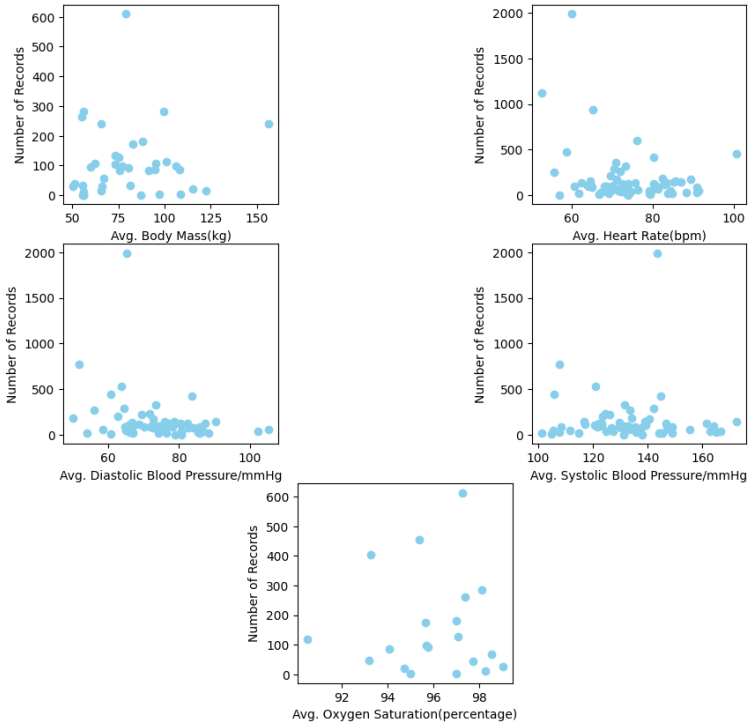


Figure 2. Scatter plot of number of observations over different health indexes: In these plots, each point represents a single patient. The x-axis is the average of all inputs of the patient. And the y-axis is the number of observation entries the patient made.

We plot the population density along different number of entries (Figure 3). From these pictures we could see that most patients having fewer inputs, which means we could be adding more notifications for patients to record.

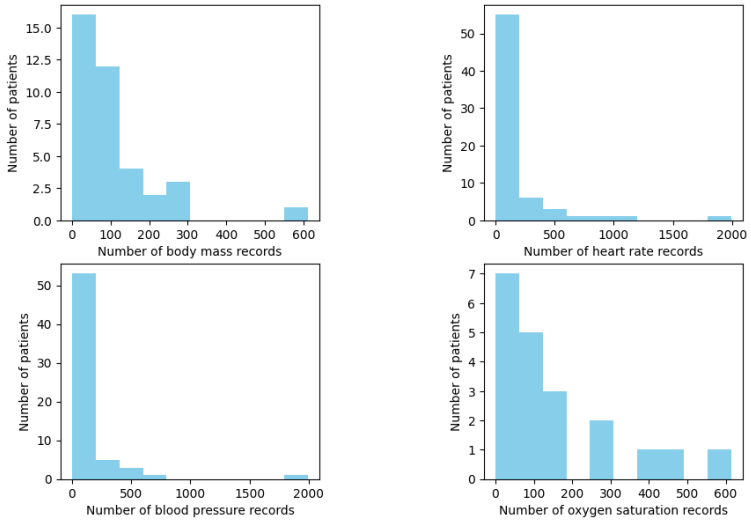


Figure 3. The histograms for the population distribution with different number of observation entries.

After exploring the Observation data set, we combine the Observations and the Demographics datasets. By analyzing the number of observation entries from different age groups by using the demographics data set, we found that male patients monitored their observation on Vironix platform more than the female patients (Figure 4).

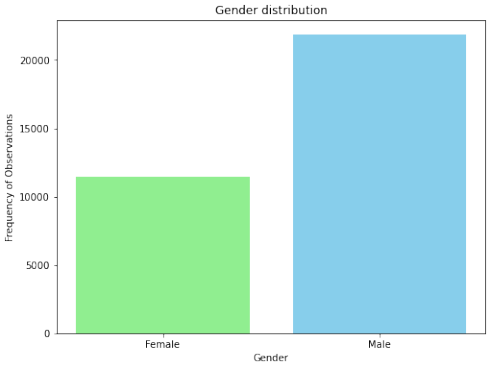


Figure 4. Frequency of using the Vironix platform by gender distribution

We studied the same frequency with respect to the age distribution of the patients and observed that individuals between 75 to 85 years old recorded the most observations. (Figure 5)

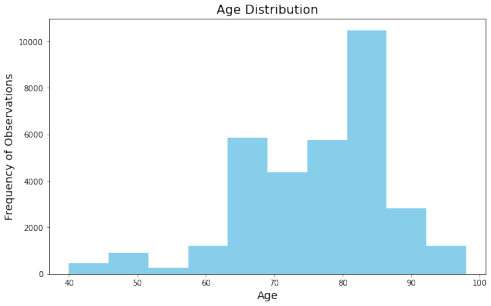
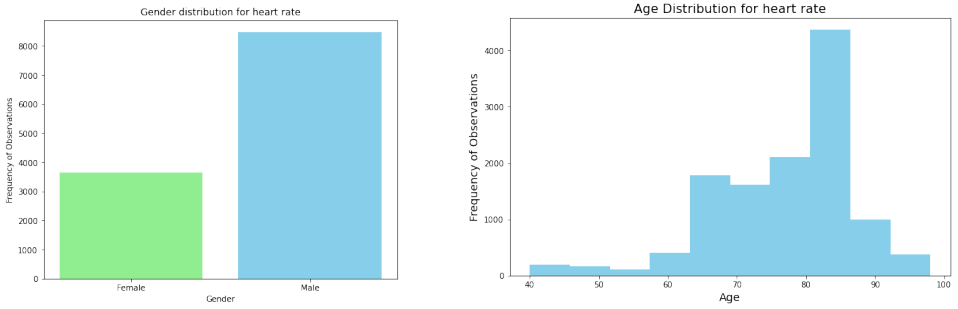


Figure 5. Frequency of using the Vironix platform by age distribution

We wanted to understand if similar trends show when we make these comparisons with only a specific vital monitoring instead of all vitals. We found that the trend indeed remains the same as seen in the following graph where we compare the frequency of noting heart rate with respect to gender and age distributions respectively (Figure 6).



(a) Gender distribution for heart rate monitoring

(b) Age distribution for heart rate monitoring

Figure 6. Frequency of recording heart rate by demographics distribution

3.2 Encounters

The encounter dataset contains information about encounter class (inpatient, outpatient, virtual, emergency and unknown type of data) and encounter types (virtual telehealth check, follow-up visit, emergency, routine checkup, hospital admission, specialty consultation, etc.). It also has information about the start date of the encounter which is collected from the medical history of the patient. We assumed that for all the patients, they started using the Vironix platform from 2021-01-01.

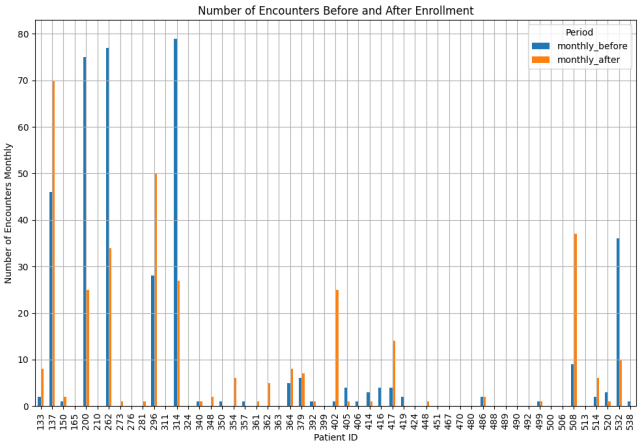


Figure 7. Monthly Frequency of Encounters Before and After Enrollment with the Vironix System

We created a dataframe about number of encounters of patients monthly and as well as yearly before they enrolled in Vironix system and after they enrolled in Vironix system. From figure 7, we can see monthly frequencies of encounters for each patient before and after enrollment on the Vironix platform. However, we weren't able to conclude the exact pattern as we see lots of variation. It happened due to shortage of data for each different type of encounter. In the sub-dataset representing encounters before using Vironix, more than 90% data was unknown about the type of encounters.

We wish to see what type of patients engage with the clinical staff most often. From figure 8, we can conclude that males around the age group of 51-60 engaged more with the clinical staff and females around the age group of 71-80 engaged more with the clinical staff.

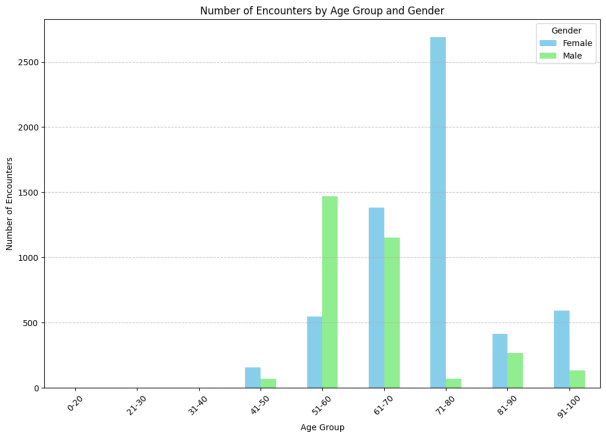


Figure 8. Number of encounters with clinical staff by age group and gender

From figure 9, we conclude that females engage more with the clinical staff than males.

Also, the average age for female interacting more with clinical staff is 71 and average age for male interacting more with clinical staff is 62.

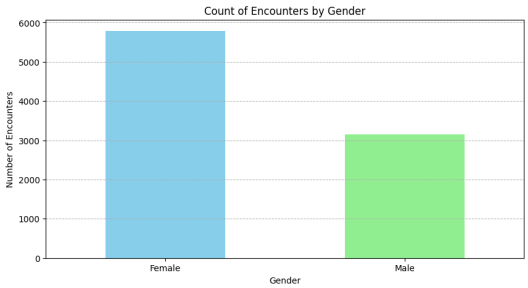


Figure 9. Genderwise comparison of number of encounters with the clinical staff

We wanted to study how these demographics impact the number of medical encounters patients have after joining Vironix but observed that we do not have enough data to make significant conclusions about it. We suggest collecting data for a longer period of time regarding the patient's encounters and the types of encounters (like outpatient, ER, inpatient, hospitalization, etc.).

3.3 Using Machine Learning and Clustering Algorithm for Early Detection and Prevention

3.3.1 *Multiple Linear Regression Model*

Using the multiple linear regression analysis, we find the function to predict patient's average heart rate based on their other characteristics (smoking habits, age, gender, height, weight and average heart rate): $f(\text{average heart rate}) = 1.81 \cdot \text{smoking habit} - 0.092 \cdot \text{age} + 1.80 \cdot \text{gender} + 0.036 \cdot \text{height} + 0.054 \cdot \text{weight}$. Figure 10 shows the actual average heart rate versus the predicted average heart rate for the test data.

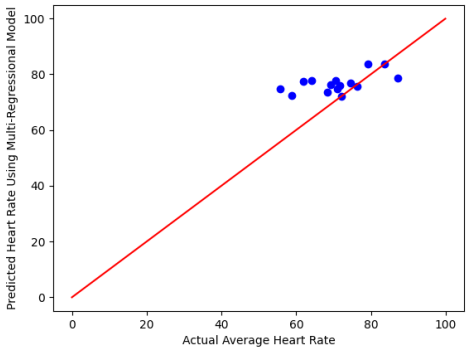


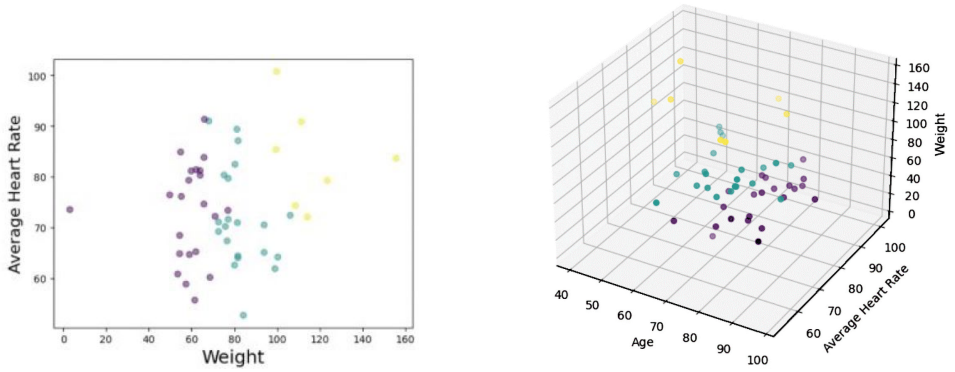
Figure 10. Actual average heart rate versus the predicted average heart rate

The coefficient of determination $R^2 = -0.17$ which implies a poor model fit or inap-

propriate features. We suggest that the above characteristics including the average heart rate should be used to predict the number of encounters of patients with the hospitals when more such data is made available. It would be more useful and practical that a multiple linear regression model is used to predict the number of encounters of patients with the hospitals based on some measurable patients characteristics. It may accomplish Vironix's goal to reduce the number of hospitalizations.

3.3.2 *Clustering Algorithm (K-Means Clustering)*

We apply K-Means Clustering algorithm to classify patients based on six characteristics (smoking habits, age, gender, height, weight and average heart rate) of the patients. Figures below show a 2D and 3D plot of patients being divided into three clusters, respectively. We can see how the three clusters relate to the three characteristics age, average heart rate and weight. Cluster 1 (in yellow dots) mostly consist of patients who are overweight and have high average heart rates. Cluster 2 (in green dots) mostly consist of patients who have high average heart rate. Cluster 3 (in purple dots) mostly consist of patients who are older and medium to high average heart rate. This result may suggest that patients in clusters 1, 2 and 3 are the patients who should be monitored frequently, moderately and occasionally, respectively. Also, by identifying which patients may need to be monitored closely can help achieve the goal of early detection and reduction in hospitalization events.



(a) 2D plot of three clusters of patients (b) 3D plot of three clusters of patients

Figure 11. K-means clustering on patient characteristics

4 Modeling degradation of CKD using GFR

The dataset used for this modeling is the Chronic Kidney Disease Research of Outcomes in Treatment and Epidemiology (CKD-ROUTE). It is an observational cohort study of a representative Japanese population with stage G2±G5 chronic kidney disease according to the Kidney Disease Improving Global Outcomes (KDIGO) classification, excluding patients undergoing dialysis. Over 1,000 participants were enrolled at the Tokyo Medical and Dental University Hospital and its 15 affiliated hospitals in the Tokyo metropolitan area of Japan. Details of the study can be found in [5].

The dataset consists of 1,138 observations with 51 variables. Some of the key features included in the model are: eGFR: measured at six different time periods (0M, 6M, 12M, 18M, 24M, 30M, 36M), Serum Creatinine Level, Serum Albumin Level, Hemoglobin Level, Body Mass Index (BMI), Blood Pressure Level, Demographics etc.

Degradation of kidney function is a process that is often described using an estimate of its Glomerular Filtration Rate (eGFR). This is a quantity that this measured in the dataset found from [5]. We would like to find statistical methods to observing the change in this quantity across the three years the data was gathered. In particular, we want to understand how important and easy to access features of a patient change the kidney malfunction progression. To this end we came up with four main methods or models. These are listed below:

(1) **Gaussian Process Regression:** Here we take a parameter-free Bayesian regression approach called Gaussian Process (GP) Regression. A GP is a collection of random variables, any subset of which has a joint Gaussian distribution. In the context of regression, a GP defines a distribution over functions, where each point in the domain has a mean and standard deviation that follow a Gaussian distribution.

(2) **Deep Learning Model**: Here we use three layers of neural networks with a training dataset and a mean square error loss function to predict eGFR values from initial time eGFR values, demographic variables, and health-related variables. The results here showed the most promise.

(3) **Classical Decay Model:** Here we take a more agent based approach. Namely, we assume that the filtration rate is based on a set of "filters" in the kidney and that this filters degrade in time loosing this capacity to filter. In this simple model we also assume that the filters are not repaired or replaced as they wear out. With this understanding we get the following equation for the eGFR.

(4) **Chain Decay Model:** In this case we bring in the fact that kidney disease progresses in multiple stages and that the filtration rate has a different relationship with features at different stages. We also add in the possibility of repair/replacement for the aforementioned filters.

The later two parameterized models are verified using curve fitting of the data from the [5] paper.

4.1 Predicting eGFR Decay Using Gaussian Process Regression

Gaussian Process Regression is a fairly simple parametric-free Bayesian regression approach with a few distinct advantages. The parametric-free nature hopefully allows the model to capture the complex interactions between a patient's vitals and their expected kidney function over the next couple of months. Furthermore, the probabilistic nature of the model allows for quantification of uncertainty, which is crucial for risk-adverse decisions often seen in a clinical setting. This modeling approach was applied to two separate predictive outcomes to kidney health: eGFR six months from now, and the deterioration rate of eGFR over the next thirty-six months. Both metrics for future kidney function can hopefully aide in informed decisions.

4.1.1 *Modeling eGFR in Six Months*

One of the most basic questions that can be asked is what is the eGFR rate a few months into the future. Given the dataset that was provided, we chose to only predict what the eGFR rate was six months out. Thus we trained our model on what the patient features and their eGFR levels when they first appear at the hospital. Figure 12 shows the results.

Unfortunately, the results of using GPR to predict this rate did not get results close to what would be useful within a clinical setting. A better kernel would vastly improve its performance. Despite this, GPR does start to approximate the distribution of results that you would expect to see.

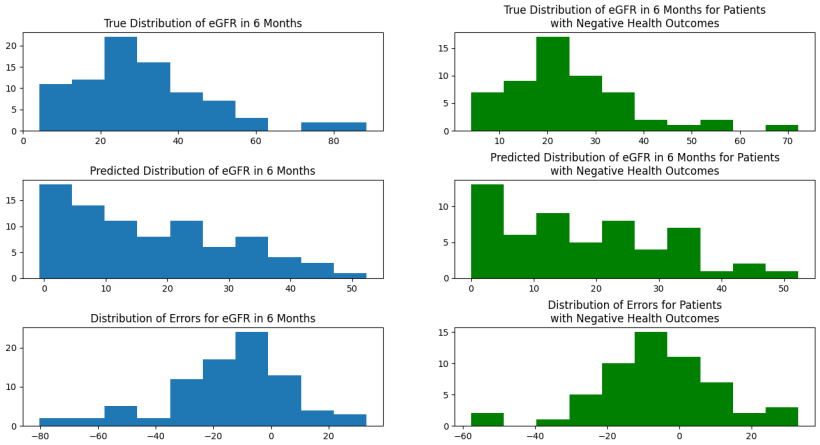


Figure 12. A comparison of the true versus predicted distribution by using the GPR model to predict patient eGFR 6 months out.

4.1.2 *Modeling eGFR Deterioration Rate*

One metric that clinicians might be interested in, is the rate in which the patient deteriorates over the next period of time. Given the slow acting nature of chronic diseases, controlling deterioration rates can be an effective way of preventing negative health outcomes in patients. When a patient has their eGFR measure, the clinician will need to decide if any intervening action should be taken. The deterioration rate should capture all the information the physician should need to make this decision.

In order to train out GPR model, we need to calculate the deterioration rate of patients over the next thirty-six months. Naively if eGFR_0 is the initial eGFR of the patient, and eGFR_{36} is the eGFR of the patient thirty-six months later, then the deterioration rate could be given by $\frac{\text{eGFR}_{36}-\text{eGFR}_0}{36}$. However given the noisy nature of both the nature of eGFR measurements and the long list of confounding variables potentially affecting eGFR, this approach is likely to be inaccurate. Instead we calculate deterioration rate d_{36} to be the slope of the line of best fit that passes through the initial eGFR and best approximates eGFR measurements over the next thirty-six months. This can be

calculated through a standard least squares approach:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 6 & 1 \\ 12 & 1 \\ \vdots & \vdots \\ 36 & 1 \end{bmatrix}$$

$$\vec{b} = [\text{eGFR}_0, \text{eGFR}_6, \text{eGFR}_{12}, \ldots, \text{eGFR}_{36}]^\top$$

$$\vec{x} = [d_{36}, \text{eGFR}_0]^\top$$

$$\vec{x} = \text{argmin}_{\vec{x}} \|\mathbf{A}\vec{x} - \vec{b}\|_2$$

After calculating the deterioration rate for each patient, we can then use this as the target variable for our GPR model. We can then use the other features of the patient to predict the deterioration rate of the patient.

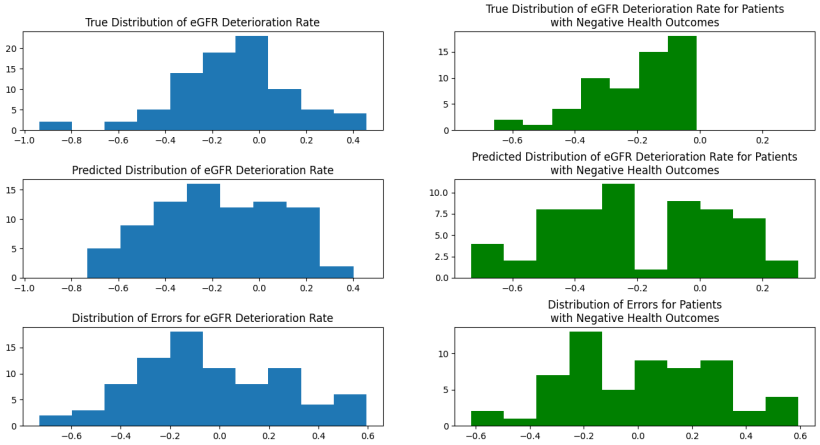


Figure 13. A comparison of the true versus predicted distribution by using the GPR model to predict patient eGFR deterioration rate. Of note is the tendency for the model to predict positive health outcomes when it should not.

The results of such a model are given in the file **gpr.ipynb** and in figure 13. The model does not perform as well as just predicting 36 months out. Furthermore, the model tends to predict that patients won't deteriorate as much as they actually do. The second column in figure FIGURE exemplifies this. Perhaps finding a better kernel for the model will improve results. A GPR model is likely not the best model for this task, however we believe that predicting deterioration rate rather than just eGFR thirty-six months out is a useful metric.

4.1.3 *Future Works*

Despite our best efforts, the limited time allocated for this approach encroached on the true effectiveness of this approach. Not enough time was given for selecting a proper kernel / covariance matrix. As of right now, the standard deviation predictions are so large that they aren't nearly as useful as they could be. Incorporating proper priors into the GRP models would also be beneficial.

4.2 Predicting eGFR Decay Using PCA and Classical Decay Models

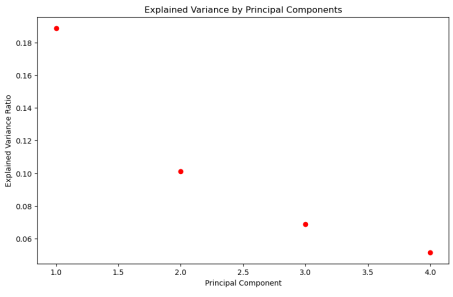
The objective of this submodel is to predict the decay of eGFR at various time periods in patients with CKD. We began the modeling process by preparing and cleaning the dataset. This involved handling missing values, normalizing the data, and encoding categorical variables using one-hot encoding.

4.2.1 *Principal Component Analysis*

To reduce the dimensionality of the dataset and identify the most significant variables, we performed Principal Component Analysis (PCA). This step was crucial in simplifying the model without losing important information. The results from the PCA indicated that the first principal component, which included various eGFR measurements, captured a significant portion of the variance in the dataset. The explained variance by each principal component is visualized in the scree plot below:

	PC1	PC2	PC3	PC4
age	-0.059105	-0.158466	-0.057784	0.014922
SBP	-0.021517	0.090176	-0.090455	-0.013455
BMI	-0.046206	0.025776	-0.164993	0.021794
Hb	0.142398	0.063381	-0.137592	-0.085289
Alb	0.081716	-0.161591	0.045123	-0.169917
Cr	-0.225580	-0.059740	0.169716	-0.077296
UPCR	-0.074825	0.236259	-0.051573	0.203696
eGFR(0M)	0.262532	0.129977	-0.087609	0.040105
eGFR(6M)	0.269172	0.106740	-0.092514	0.014963
eGFR(12M)	0.270752	0.097607	-0.083450	-0.011930
eGFR(18M)	0.271166	0.080181	-0.077024	0.011196
eGFR(24M)	0.272832	0.075660	-0.080702	-0.000095
eGFR(30M)	0.272994	0.061074	-0.072926	-0.008348
eGFR(36M)	0.270799	0.054922	-0.067294	0.007997
gender_1	-0.014376	-0.032961	-0.131780	-0.197872
gender_2	0.014376	0.032961	0.131780	0.197872
etiology of CKD_1	-0.111784	0.084156	-0.092029	0.243663
etiology of CKD_2	-0.010691	-0.205368	-0.125948	-0.149742
etiology of CKD_3	0.066982	0.209339	0.056768	-0.101371
etiology of CKD_4	0.057285	-0.039521	0.203549	0.059884

(a) PCA result



(b) scree plot

The PCA analysis successfully identified key features contributing to eGFR in CKD patients. it also helped in reducing the number of variables while retaining essential information. The significant contribution of eGFR measurements and proteinuria-related variables highlights their importance in monitoring CKD progression.

4.2.2 *Classical Decay Prediction*

We utilized the Minimal Decay Model to predict eGFR values. This model is defined by the equation:

$$X(t) = X_0 \exp\{-w^T f\}t$$

where

- $X(t)$ is the eGFR at time t.
- X_0 is the initial eGFR
- w is a vector of coefficients
- f is a vector of features and t is time

In log form:

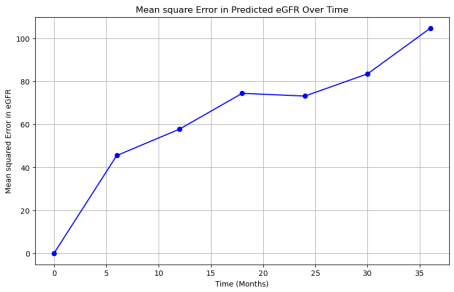
$$\log(X(t)) = \log(X_o) - (w^T f)t$$

The data was fitted to this model using the following features to predict the decay of eGFR: Age, Systolic blood pressur, Body mass index, Hemoglobin level, albumin level, creatinine level.

The predictions were made for eGFR values at six-month intervals over a 36-month period. Below is a comparison of the actual vs. predicted eGFR values for the **first two patients**:

	Time	Actual	Predicted
0	0	34.146986	34.146986
1	6	26.454698	33.903286
2	12	24.331582	33.661325
3	18	24.682189	33.421090
4	24	21.614854	33.182570
5	30	20.420524	32.945753
6	36	18.495328	32.710625
7	0	73.570568	73.570568
8	6	78.287758	72.720524
9	12	71.343858	71.880302
10	18	72.845992	71.049787
11	24	71.908942	70.228869
12	30	71.562914	69.417435
13	36	67.225032	68.615377

(a) Model result



(b) Mean Square Error

The mean squared error (MSE) between the actual and predicted eGFR values over time was calculated to evaluate the model's performance.The model starts with a relatively low error at the initial time point (0 months), indicating that the initial conditions were well-captured by the model.As time progresses, the error increases, which is evident from the upward trend in the MSE plot. This suggests that the model's predictions become less accurate over longer time intervals.

4.3 Predicting standardized eGFR using Deep Neural Network

We want to use the standardized eGFR as the predicted outcome of the network:

$$\hat{eGFR} = \frac{eGFR}{\text{mean}(eGFR)}$$

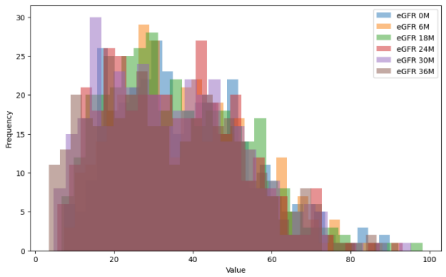


Figure 16. Histogram of eGFR

We expect to predict the standard eGFR after 6, 12, 18, 24, 30, 36 months using the current eGFR (at 0 month) and some demographic and health-index variables: age, gender, SBP, BMI, proteinuria. The procedure can be extended using more explanatory variables. The description of eGFRs is in figure 16.

Network architecture: Fully connected neural network was used with 3 layers: 10, 20, 10 neurons in each layer. We used tanh activation function to capture both negative and positive values of standardised eGFR. The input are explanatory variables, and the output are 6 predicted values of eGFR. Loss function is mean square error type, combining component losses from each part of the output.

Data: After skipping NA observations, we used data from 441 follow-up patients, split 80% in train set and 20% in test set.

Results:

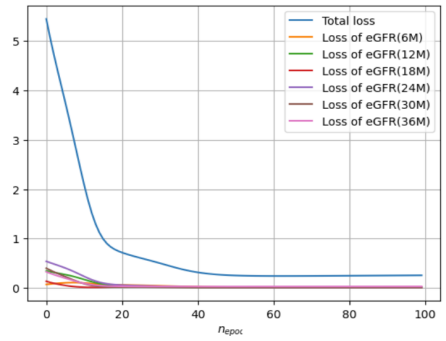
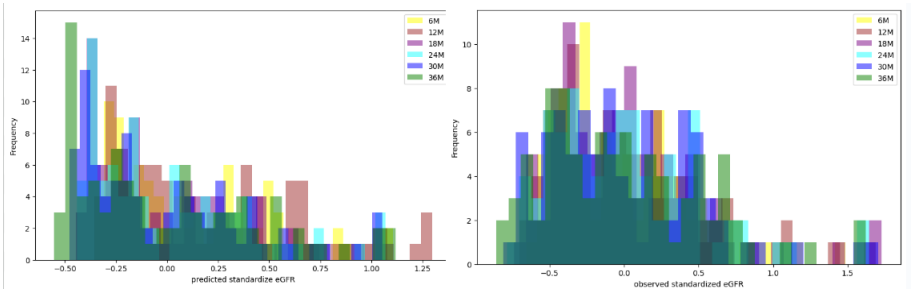
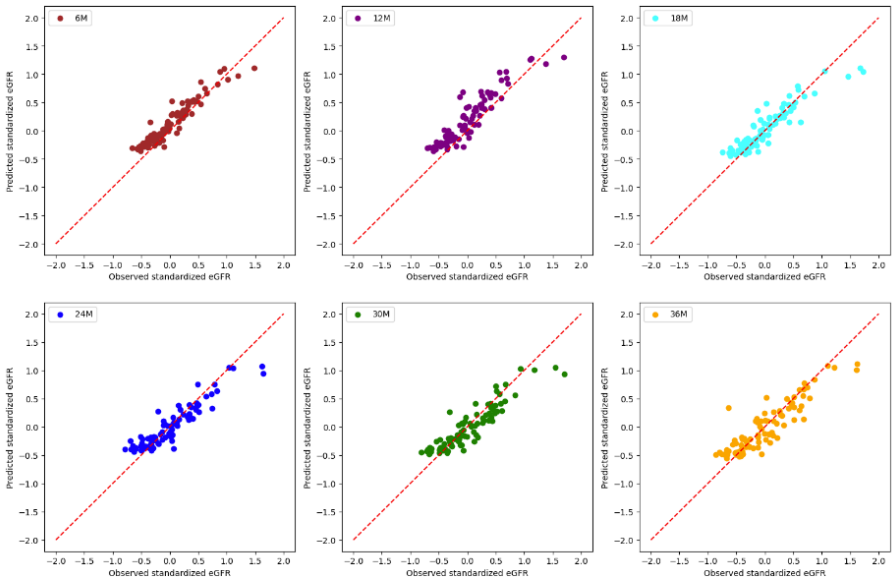


Figure 17. Loss function and its components.



(a)



(b)

Figure 18. Predicted vs Observed standardised eGFR.

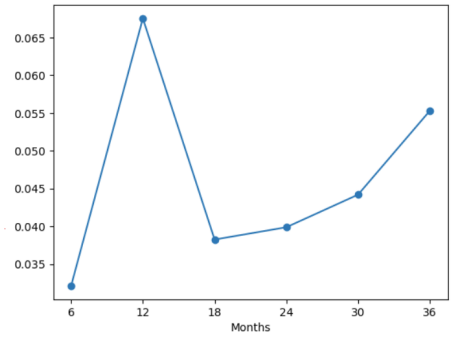


Figure 19. Mean square error of standardised eGFR estimation along the time.

Figure 17 shows good convergence of the component losses on the train set after 100 iterations. On test set, figure 18 reveals a light difference in the distribution of predicted

eGFRs versus the observed ones, as well as provides a closer look at the prediction of single parts in the output, raising a question about under-estimation of future eGFRs given the current one. The mean square error in figure 19 shows the overall increasing trend over time, except for the case predicting eGFR after 12 months.

4.4 Predicting eGFR Decay Using Chain Decay Model

We are looking at the glomerular filtration happening due to units we refer to as filters. In this submodel, these filters are working at different performance levels as the CKD progresses through its stages. The number of filters performing at stage i is given by x_i, and the filtration rate at stage i is given by v_i. Hence, the total filtration is

$$\frac{d\hat{F}}{dt} = \sum_{i=1}^{n} x_i v_i$$

Where $v_i > v_j$ for each $i < j$. Now we look at the progress of each filter population x_i. We expect that the features affecting population decline have a different relationship and significance at different stages. We also expect that some of the filter populations have a renewal rate where more filters are introduced. It seems intuitive this would only occur for stage 1 filters (x_1). For the sake of generality, we will ascribe a renewal rate term for each stage. These considerations are captured by the following system of equations.

$$\begin{aligned}\frac{dx_1}{dt} &= -\lambda_1 x_1 + r_1 \\ \frac{dx_i}{dt} &= -\lambda_i x_i + \lambda_{i-1} x_{i-1} + r_i\end{aligned}$$

Where λ_i are population decay rates. The λ_i values will be given by dot products of features chosen and their respective weights. Solving this system we get the following solution.

$$x_i(t) = s_i + \sum_{k=1}^{i} d_{ik} e^{-\lambda_k t}$$

Where d_i and s_i are constant coefficients that we determine using initial conditions and the decay parameters (λ_i). Putting this back into the total filtration rate we see that we get a linear combination or mixture of exponential functions and a constant modeling the filtration rate. This is given by

$$\frac{d\hat{F}}{dt} = \sum_{j=1}^{n} \sum_{i=1}^{j} s_i v_j + d_{ji} v_j e^{-\lambda_i t}$$

For simplicity, we take $v_j = \frac{f_0}{x_1(0)} \frac{\text{unit}}{\text{sec}}$ where f_0 is the initial total filtration rate. This implies the the filter populations are currently only different in their life span but not in their filtration performance. We also take $r_i = 0$ for all i greater than one. Next, we normalize the filtration rate by the equation $\frac{dF}{dt} = \frac{d\hat{F}}{dt} / f_0$. Then we get

$$\frac{dF}{dt} = \frac{1}{x_1(0)} \sum_{j=1}^{n} \sum_{i=1}^{j} s_i + d_{ji} e^{-\lambda_i t}$$

For our implementation we will only have two stages unlike the previously mentioned five stages. The reason for using a two-stage chain decay model is because we only have seven time points for each patient and, consequently, too few data points for a many-chain model's regression. In other words, we run the risk of over-fitting if we take a significant number of stages. With two stages we get the following coefficients:

$$
\begin{aligned}
d_{11} &= x_1(0) - (r_1/\lambda_1) \\
d_{21} &= -\frac{(x_1(0) - (r_1/\lambda_1)))\lambda_2}{\lambda_1\lambda_2 - \lambda_2^2} \\
d_{22} &= \frac{(x_1(0) - (r_1/\lambda_1)))\lambda_2}{\lambda_1\lambda_2 - \lambda_2^2} - \frac{r_1/\lambda_1}{\lambda_1\lambda_2 - \lambda_2^2}(\lambda_1 - \lambda_2) \\
s_1 &= r_1/\lambda_1 \\
s_2 &= \frac{r_1/\lambda_1}{\lambda_1\lambda_2 - \lambda_2^2}(\lambda_1 - \lambda_2)
\end{aligned}
$$

and we get the filtration rate

$$
\frac{dF}{dt} = \frac{1}{x_1(0)}(s_1 + d_{11}e^{-\lambda_1 t} + s_2 + d_{21}e^{-\lambda_1 t} + d_{22}e^{-\lambda_2 t})
$$

(We notice that this solution is somewhat similar to what is called Bateman's equation.) We arbitrarily pick $x_1(0)$ to be 100. Initially, we pick one feature and observe in which stage it has more significance. The parameters will be defined as $\lambda_1 = w_1 f_1, \quad \lambda_2 = w_2 f_1, \quad r_1 = w_3 f_1$. Here f_1 is the one feature we use in this initial study. In our case that is BMI. w_1 to w_3 are the weights. We initialize w_1 and w_2 with increasing magnitude as we increase the stages since we expect higher stages have faster degradation. It would be ideal to have a linear combination with more features for each decay rate but we will prioritize parameter minimization and simplicity for now. Figure 20 shows the fitted normalized eGFR curve for two randomly picked patients across our time period. The eGSR values are normalized by the initial (0 Month) eGSR value.

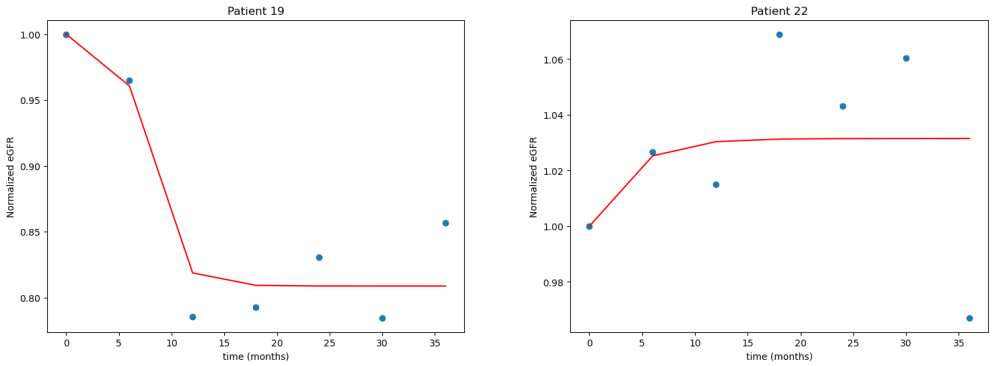


Figure 20. Sample trends of normalized eGFR with Body Mass Index (BMI) as our feature.

The mean square for the first 407 patients is given in figure 21. We only consider 407 of

the total 417 (from the cleaned dataset) because of rare outliers for whom the nonlinear regression does not converge.

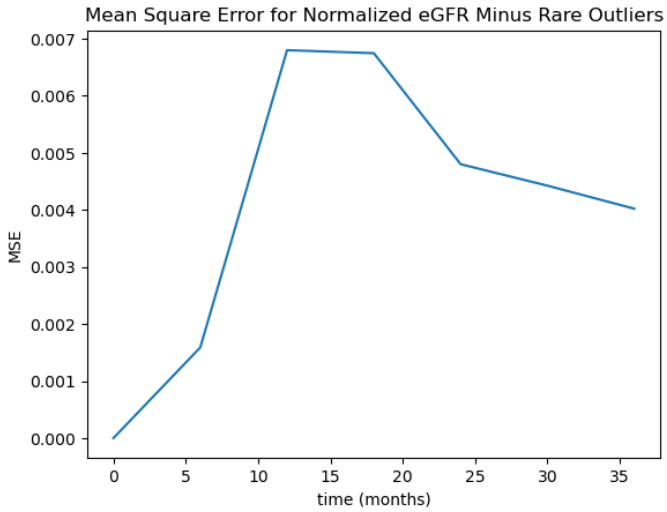


Figure 21. Mean square error of chain (multistage) decay model's regression for 407 patients.

5 Future Directions

Future work include analyzing the changes in patient's symptoms which will provide further information for patient's preventive care and reduction in hospitalization events, predicting the number of encounters of each patient in hospitalization, utilizing the survey questions to further improve patient's experience and level of satisfaction with Vironix.

For kidney degradation modeling we would be interested in obtaining more data both for our machine learning approaches and our phenomenological models. In particular, we would be keen to get time series data on the evolution of the various features of the patients such as BMI. This would help us make better predictions on future eGFR models. We we would also like to investigate the efficacy of the eGFR as a measure of Glomerular Filtration Rate. Finally, we would like to account for circumstances we don't comprehend and/or measure in kidney degradation by adding stochasticity to our phenomenological models.

References

[1] *https://www.cdc.gov/kidneydisease/pdf/chronic-kidney-disease-in-the-us-2021-h.pdf*. Accessed: 2024-06-27.

[2] *National institute of diabetes and digestive and kidney diseases*. Accessed: 2024-06-27.

[3] Bai, Q., Su, C., Tang, W., & Li, Y. (2022). Machine learning to predict end stage kidney disease in chronic kidney disease. *Scientific reports*, **12**(1), 8377.

[4] Dritsas, E., & Trigka, M. (2022). Machine learning techniques for chronic kidney disease risk prediction. *Big data and cognitive computing*, **6**(3), 98.

[5] Iimori, S., Naito, S., Noda, Y., Sato, H., Nomura, N., Sohara, E., Okado, T., Sasaki, S., Uchida, S., & Rai, T. (2018). Prognosis of chronic kidney disease with normal-range proteinuria: The ckd-route study. *Plos one*, **13**.

[6] Inker, L.A., Eneanya, N.D., Coresh, J., Tighiouart, H., Wang, D., Sang, Y., Crews, D.C., Doria, A., Estrella, M.M., Froissart, M., & Grams, M.E. (2021). New creatinine- and cystatin c–based equations to estimate gfr without race. *The new england journal of medicine*, **385**(19), 1737–1749.

[7] Major, R.W., Shepherd, D., Medcalf, J.F., Xu, G., Gray, L.J., & Brunskill, N.J. (2019). The kidney failure risk equation for prediction of end stage renal disease in uk primary care: an external validation and clinical impact projection cohort study. *Plos medicine*, **16**(11), e1002955.

[8] Mota-Zamorano, S., Robles, N.R., Lopez-Gomez, J., Cancho, B., González, L.M., Garcia-Pino, G., Navarro-Pérez, M.L., & Gervasini, G. (2021). Plasma and urinary concentrations of arachidonic acid-derived eicosanoids are associated with diabetic kidney disease. *Excli journal*, **20**, 698.

[9] Stevens, L.A., Coresh, J., Greene, T., & Levey, A.S. (2006). Assessing kidney function — measured and estimated glomerular filtration rate. *The new england journal of medicine*, **354**(23), 2473–2483.

[10] Tangri, N., Stevens, L.A., Griffith, J., Tighiouart, H., Djurdjev, O., Naimark, D., Levin, A., & Levey, A.S. (2011). A predictive model for progression of chronic kidney disease to kidney failure. *Jama*, **305**(15), 1553–1559.

[11] Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., Zhu, S., & Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine*, **17**, 1–13.

[12] Ye, H., Chen, Y., Ye, P., Zhang, Y., Liu, X., Xiao, G., Zhang, Z., Kong, Y., & Liang, G. (2021). Nomogram predicting the risk of three-year chronic kidney disease adverse outcomes among east asian patients with ckd. *Bmc nephrology*, **22**, 1–9.