

# **Optimal Advanced Chronic Kidney Disease Phenotyping using Combined Massachusetts General Brigham Electronic Healthcare Records and Medicare Insurance Claims Data - Generalized to the Open Source MIMIC-III Clinical Database**

Lily Bessette, Lukas Katko, Nathaniel Wilson  
bessette.l@northeastern.edu | katko.t@northeastern.edu | wilson.na@northeastern.edu

Khoury College of Computer and Information Science  
Northeastern University, Boston, MA  
December 14, 2020

## **1. Objectives and significance**

The goal of this project is to build and compare three different prediction models for identifying clinically important phenotypes. The models used will be Logistic Regression, LASSO Regression, Random Forest, and Deep Neural Networks trained on patient data and serum creatinine laboratory values in order to predict whether a given patient is likely to have advanced chronic kidney disease (CKD). In this project, a laboratory result of serum creatinine will be used to compute a patient's estimated glomerular filtration rate (eGFR) in order to define advanced chronic kidney disease. These lab value defined target labels used in our prediction models are hard clinical endpoints that ultimately inform physicians of the level of kidney function in their patients and therefore the proper methods of care for patients likely to have advanced chronic kidney disease.

About 37 million people in the United States, about 15% of the US population, are estimated to have chronic kidney disease. According to the CDC, people may not feel ill or notice any symptoms until CKD is advanced and most adults (9 in 10) with CKD do not know they have it. Furthermore, almost half of the people with very low kidney function who are not on dialysis do not know they have CKD. CKD related health problems also involve early death, higher risk of heart disease and stroke, and kidney failure/end stage renal disease [15]. Therefore the ability to find predictors for advanced CKD in a patient population is important for prevention, which from a physician perspective can involve more closely monitoring patients with these predictors and additional prompts to test for CKD. The tests for CKD are through simple blood and urine tests. Of interest, the blood test for creatinine, which is a waste product produced by the muscles, is the measure that will be studied, and, by measure of eGFR, be predicting kidney function. Additionally, the predictors found in this project can enable further pharmacoepidemiological comparative effectiveness research in adding these predictors to their variable

selection for confounding adjustments and to their propensity-score models to balance patient populations when comparing prescription medication use in CKD patients.

Overall, a higher classification accuracy was found across all models trained and tested on the Massachusetts General Brigham (MGB) EHR-insurance claims linked dataset as opposed to those trained and tested on the Beth Israel Deaconess Medical Center (BIDMC) MIMIC-III Clinical dataset, but a lower balanced accuracy due to lower prevalence of advanced CKD in the MGB dataset. Furthermore, advanced machine learning methods such as Random Forest and Neural Networks do not perform materially better than basic modeling techniques such as Logistic and Linear Regression in classifying patients above and below an eGFR threshold of 30 for presence or absence of advanced CKD. Neural Networks showed the best results on the MIMIC-III dataset, however performed worse on the MGB dataset than the Linear/Logistic regressions models.

## **2. Background**

### **2.1 Models**

One binary classification model and three regression models were implemented. Logistic regression is a binary classifier that utilizes a sigmoid activation function to find coefficient weights that maximizes the likelihood to ultimately form a predictor of the posterior probabilities that are then converted to a predicted class output using the application of the maximum a posteriori principle. Logistic regression with RIDGE applies a penalty to the L2 Norm to reduce complexity and apply regularization in updating the gradient descent rule. Similarly, a penalty to the L1 Norm can be applied called LASSO. Least absolute shrinkage and selection operator (Lasso) regression utilizes a cost function in linear regression that reduces overfitting by L1 regularization and can lead to coefficients of 0 for ‘least important’ features.

The Random Forest model was introduced by Tin Kan Ho in the 1995 paper “Random Decision Forests” to address some of the shortcomings that arose out of using single decision trees as classifiers, which had a tendency to overfit to their data [4]. Ho’s Random Forest model addresses the overfitting problem in two significant ways. First, instead of relying on a single tree, the model generates many trees, each of which gets a ‘vote’ for the class of the input data where the forest output is the mode (classification) or average (regression) of all the tree’s results. Second, each tree in the forest does not see the full set of features, but rather a randomly selected subset.[4]. This random feature assignment protects the model from over fitting by limiting the feature exposure of each tree, while addressing the full feature set in the forest as a whole [4].

Fundamentally, an Artificial Neural Network (ANN) is a model that attempts to mimic the biological processes within an organism’s brain involved in learning its environment. In this model, each

neuron is something called a perceptron. A perceptron takes a linear combination of a series of weighted inputs and applies activation function at its output. On its own, a perceptron is extremely limited. However, when stacked together in layers (Multilayer Perceptron), advanced functions are able to be described. When more than two layers are used, this type of ANN is called a Deep Neural Network, which will be the primary model used within this project for the ANN.

Each layer of the ANN will have a number of neurons, each with an output equal to the weighted combination of the output of the previous layer's neurons. An activation function is then applied to each output. Some common activation functions are Sigmoid function:  $\sigma(z) = 1 / (1 + \exp(-z))$ , Hyperbolic Tangent function:  $\tanh(z) = 2\sigma(2z) - 1$ , or the Rectified Linear Unit function:  $\text{ReLU}(z) = \max(0, z)$ . Finally, the output of the final layer is used to make a prediction and then compared to a label or target value in order to measure the network's error using a specified loss function [9].

## 2.4 Previous Work

A similar research study was conducted in a Taiwanese insurance claims dataset that implemented a series of models: logistic regression, decision tree, random forest, XGBoost, AdaBoost, LightGBM, and convolutional neural networks (CNN). The CNN approach with a model with a 6-month feature assessment window was reported to have performed best with an AUROC of 0.957 [16]. This study differs from this project substantially since the target variable is measured by the occurrence of ICD-9 codes 585 and 586. Patients with these two codes are categorized as having CKD, whereas propensity matched control patients without these codes are categorized as not having CKD. This is a substantial difference since in order to receive these ICD-9 codes a patient must be diagnosed with CKD by a medical provider and the insurance claim for that medical visit must contain those ICD-9 codes. This is drastically different from using the laboratory value of serum creatinine directly to measure eGFR which in turn measures kidney function, which is this project's approach.

Similar to the goals of this project, but differing in the exact disease, Dr. Qionghjing Yuan et. al summarizes the work of other investigators in acute kidney injury [17]. Acute kidney injury (AKI) differs from advanced CKD and CKD in general because AKI is typically caused by a specific event, such as dehydration, blood loss, or medication use, whereas CKD is caused by a long-term underlying disease, such as hypertension or diabetes. Both AKI and CKD produce damage to the kidneys and a patient's renal function, but CKD is a more gradual or slow increase in damage to the kidneys [18].

Dr. Yuan reports the use of recurrent neural network, regression based methods, decision trees, random forest, extreme gradient boosting, support vector machine, gradient boosted trees, logistic regression, XGBoost, multivariate logistic regression, and artificial neural networks in previous literature

to predict AKI. Methods implemented in this project are clearly not distinctly unique or new, but the application to advanced chronic kidney disease evaluated using laboratory values is novel and unstudied [17]. In this disease space of advanced chronic kidney disease, this is a novel research project using a combination of laboratory values from electronic healthcare records and baseline characteristics from insurance claims data that will implement well studied machine learning methods.

### 3. Methods

#### 3.1 Data

Two datasets were used: (1) the open source MIMIC-III Clinical Database [12] available on PhysioNet comprised of patients who stayed in critical care units of the Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012 and (2) a private dataset of Medicare insurance claims data linked with electronic healthcare data of patients receiving care at a Massachusetts General Brigham (MGB) Integrated Healthcare System facility from 2007-2014 [19]. The MIMIC-III dataset was obtained through a data use agreement with the MIT Laboratory for Computational Physiology (MIT-LCP) via their PhysioNet Credentialed Health Data License on the PhysioNet portal. The private MGB dataset was obtained by DUA with the Centers for Medicare & Medicaid Services and under Institutional Review Board permission under IRB #2017P002659 titled “Improving causal analyses by incorporating clinical information from electronic health records into claims data analyses.”






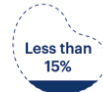
**Table 1:** Flowchart of cohort selection based on exclusion criteria

	MGB		MIMIC-III	
	<i>Less Excluded</i>	<i>Remaining</i>	<i>Less Excluded</i>	<i>Remaining</i>
All patients		569,989		46,520
Did not meet cohort entry criteria - SCr Lab Value within range (0,20]	-429,713	140,276	-7,333	39,187
Excluded due to insufficient enrollment	-53,251	87,025	NA	39,187
Excluded based on Missing Age or Gender	-10	87,015	-0	39,187
Excluded based on Dialysis	-1,272	85,743	-1,890	39,187
Final cohort		85,743		37,297

Using proper pharmacoepidemiologic study design techniques (Appendix Figure 1), the datasets (MIMIC-III and MGB) were queried for patients receiving a serum creatinine lab value (i.e. target designations) and measured the features/attributes of each patient (i.e. data point) during the baseline period of 6 months (90 days before and 90 days after) surrounding their measured serum creatinine lab value. After implementing the steps of Appendix Figure 1, the resulting number of patients for each cohort (MGB and MIMIC-III) are enumerated in the Table 1 Flowchart. The serum creatinine values found at this step were used to compute the eGFR of each patient using the CKD-EPI definition [20]:

$$eGFR = 141 \times \min(SCr/K, 1)^\alpha \times \max(SCr/K, 1)^{-1.209} \times 0.993^{Age} \times 1.018[I = female] \times 1.159[I = black]$$

(where  $K$  is .7 for females and .9 for males and  $\alpha$  is -0.329 for females and -0.411 for males). The eGFR value found from the CKD-EPI computation was used to create a binary variable where patients with an eGFR value  $< 30$  were categorized as having advanced CKD and those with an eGFR value  $\geq 30$  were categorized as not having advanced CKD (Figure 6). The continuous calculations of eGFR will be used for training/testing the Lasso regression, random forest, and neural network models. The binary interpretation of presence/absence of advanced CKD is used for training/testing the logistic regression and for evaluating the accuracy of all models after testing and conversion to binary labels.

STAGES OF CHRONIC KIDNEY DISEASE		GFR*	% OF KIDNEY FUNCTION
<b>Stage 1</b>	Kidney damage with <b>normal</b> kidney function	90 or higher	 90-100%
<b>Stage 2</b>	Kidney damage with <b>mild loss</b> of kidney function	89 to 60	 89-60%
<b>Stage 3a</b>	<b>Mild to moderate</b> loss of kidney function	59 to 45	 59-45%
<b>Stage 3b</b>	<b>Moderate to severe</b> loss of kidney function	44 to 30	 44-30%
<b>Stage 4</b>	<b>Severe</b> loss of kidney function	29 to 15	 29-15%
<b>Stage 5</b>	Kidney <b>failure</b>	Less than 15	 Less than 15%

\* Your GFR number tells you how much kidney function you have. As kidney disease gets worse, the GFR number goes down.

Figure 1 - Stages of Chronic Kidney Disease [13]

Both datasets utilize the rich electronic patient data from these two well known healthcare institutions (MGB and BIDMC). The datasets that were created for our prediction modeling contain patient features regarding patient demographics, comorbidities defined by codes from International Classification of Diseases under the Ninth Revision (ICD-9), procedures defined by CPT codes, use of prescription medications defined by National Drug Codes (NDC), a validated frailty score [11], healthcare utilization metrics, and the resulting serum creatinine laboratory values from which eGFR was calculated. The patient characteristics (features) are listed in Appendix Table C.

One of the concerning features in the MIMIC data set was patient age. Due to the need to anonymize the data, any patients older than 89 have unknown ages (for MIMIC this is represented as a random year from 1800-1900). For this reason, any patient whose age was greater than 89 was set to 91.4 - which is the median age of this population. The result of this adjustment is discussed in more detail in the results section for its impact on specific models.

## 3.2 Methods and Implementation

Initially, implementing classification techniques for presence or absence of advanced chronic kidney disease estimated from the dichotomization of the eGFR computed value was considered. Instead, a regression was performed to predict the continuous value of eGFR in linear regression with Lasso regularization, random forest, and neural network models and then dichotomize into class labels of presence of advanced CKD for an eGFR value  $< 30$  and absence of advanced CKD for an eGFR value  $\geq 30$  since the eGFR value corresponds to the percent of kidney function for a given patient and thereby the clinical stage of chronic kidney disease. In this way, the value of predicting a continuous outcome is assessed by evaluating the accuracy of Lasso regression, random forest, and neural network models in comparison to a simple logistic regression. The value of predicting a continuous outcome is clinically more meaningful than classification of presence or absence of advanced CKD since, for example, an eGFR of 31 is nearly encoded as advanced CKD, but is clinically significantly different from an eGFR of 95, normal kidney function.

For Lasso and logistic regression, the MGB dataset was subset into an 80-20 training and testing sets where a 10-fold cross validation was performed in addition to a grid search with a variation of parameters to find the best models. The feature values are standardized by centering to the mean and component wise scaled to unit variance [23, 24]. For the Lasso regression grid search,  $\alpha$ , the constant that multiplies the L1 term, was varied from 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 2, and 5. For logistic regression grid search, the penalty was varied between L1 and L2 for regularization with saga and liblinear as possible solvers to the optimization problem and C, the inverse of the regularization strength, was varied from 0.001, 0.01, 0.1, 1, 10, 100, and 1000. This process was repeated on the MIMIC-III dataset.

The Random Forest regressor was implemented using the sklearn library in Python with the goal of predicting whether a given patient has advanced chronic kidney disease from the given set of input parameters [25]. The data was split with a 70/30 training/testing. The following hyper parameters were adjusted first via 5-fold cross-validation with random search:

Hyperparameter name	Description
num_trees	Number of trees to generate in the random forest
max_depth	Maximum depth of decision tree (effectively the number of splits the data can go through before stopping)
split_samples	Number of samples which need to be in a node before it is split
min_leaf_samples	Minimum number of samples allowed in the leaf nodes
bootstrap	Whether our data sampling is done with replacement or without replacement

Once the random search identified the region of the best parameters a grid search around those values was run and selected the best model from that. Further specific details of this process are discussed in the results section. The Deep Neural Network model was trained using the TensorFlow framework. In the below table, a list of the hyperparameters that were tuned is summarized. A randomized grid search was performed to find the best selection of hyperparameters. In order to ensure that there was enough model variety, the model's depth (number of hidden layers) and width (number of neurons per layer) were also set during each phase of the randomized grid search, effectively becoming two more parameters to tune.

Hyperparameter name	Description
num_layers	Number of hidden layers
num_neurons	Number of neurons per layer
activation	Activation applied to each neuron's output
dropout	Percentage of each layer's neurons to turn off during each phase of training
optimizer	Optimization technique
lr	Learning rate of the optimizer

Hyperparameters considered for DNN model optimization

At the outset of training, 20% of the dataset was set aside as the test set. The remaining 80% was then split further into a training set and a validation set with a ratio of 80% to 20%. The validation set was used to assess each of the randomized models created by the grid search, as well as to monitor during training to prevent overfitting. If the validation data's loss did not improve for a fraction of the total epochs, the model stopped training.

With the best parameters identified, a new model was trained across all of the training and validation data together. For all versions of the model, Mean Absolute Error was used as the loss function. Typically, Mean Squared Error is used as the loss function in Neural Network regression problems, however in some cases the model would diverge during training. The results of this model with the test set are shown below in the Results section.

### 3.3 Evaluation

Since the models were trained to output a predicted continuous eGFR value through regression, the performance of the regression models was evaluated by comparing the mean squared error (MSE) defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where  $\hat{y}_i$  is the predicted value output from the model and  $y_i$  is the actual value. The difference is squared, summed over all points and then divided by the number of points. This summarizes the difference in model eGFR predictions and actual eGFR values for each model in each data set. The final MSE values were calculated using the test set that was set aside prior to the model training. The logistic regression models were evaluated on balanced accuracy, ROC AUC, Precision-Recall AUC, and Matthews correlation coefficient. The ROC curve plots the true positive rate vs the false positive rate - whose performance can be quantified using the area under the curve to compare models. Finally, the precision-recall curve plots precision vs recall - when one model's curve dominates another, then the dominant curve's model is superior.

For all models, the predicted eGFR was used to categorize predictions and the test set into class 1 for presence of advanced CKD with an eGFR of <30 and class 0 for absence of advanced CKD with an eGFR of  $\geq 30$ . This allows us to then evaluate the classification accuracy and related classifier evaluation metrics for all models. The following standard classification evaluation metrics computed from the confusion matrix was used to evaluate the models as classifiers.

Confusion Matrix [21]		
	Predicted True	Predicted False
Actual True	True positive (tp)	False negative (fn)
Actual False	False positive (fp)	True negative (tn)

Name	Symbol	Definition
Classification Error	error	$\frac{fp+fn}{tp+fp+tn+fn}$
Classification Accuracy	accuracy	$1 - error$
True Positive Rate, Recall	tpr, rc, sensitivity	$\frac{tp}{tp+fn}$
False Negative Rate	fnr	$\frac{fn}{tp+fn}$
True Negative Rate	tnr, specificity	$\frac{tn}{tn+fp}$
False Positive Rate	fpr	$\frac{fp}{tn+fp}$
Precision	pr, PPV	$\frac{tp}{tp+fp}$
Negative predictive value	NPV	$\frac{tn}{tn+fn}$
Balanced accuracy	-	$\frac{tpr+tnr}{2}$



Harmonic Mean of Precision and Recall	F-measure, balanced F-score, $F_1$	$\frac{2 \cdot pr \cdot rc}{pr + rc}$
Matthews correlation	$\text{Corr}[f(X), Y]$ , mcc	$\frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$

## 4. Results

### 4.1 Logistic Regression

Using a grid search to choose between L1 or L2 regularization, the respective inverse of regularization strength, and solver function, the model with the best mean cross-validated score of all combinations of hyperparameter variation for each dataset was found. The model parameters selected from this grid search (below) was used to then test our resulting best logistic regression model for binary classification and compute the following evaluation metrics from the Confusion Matrices.

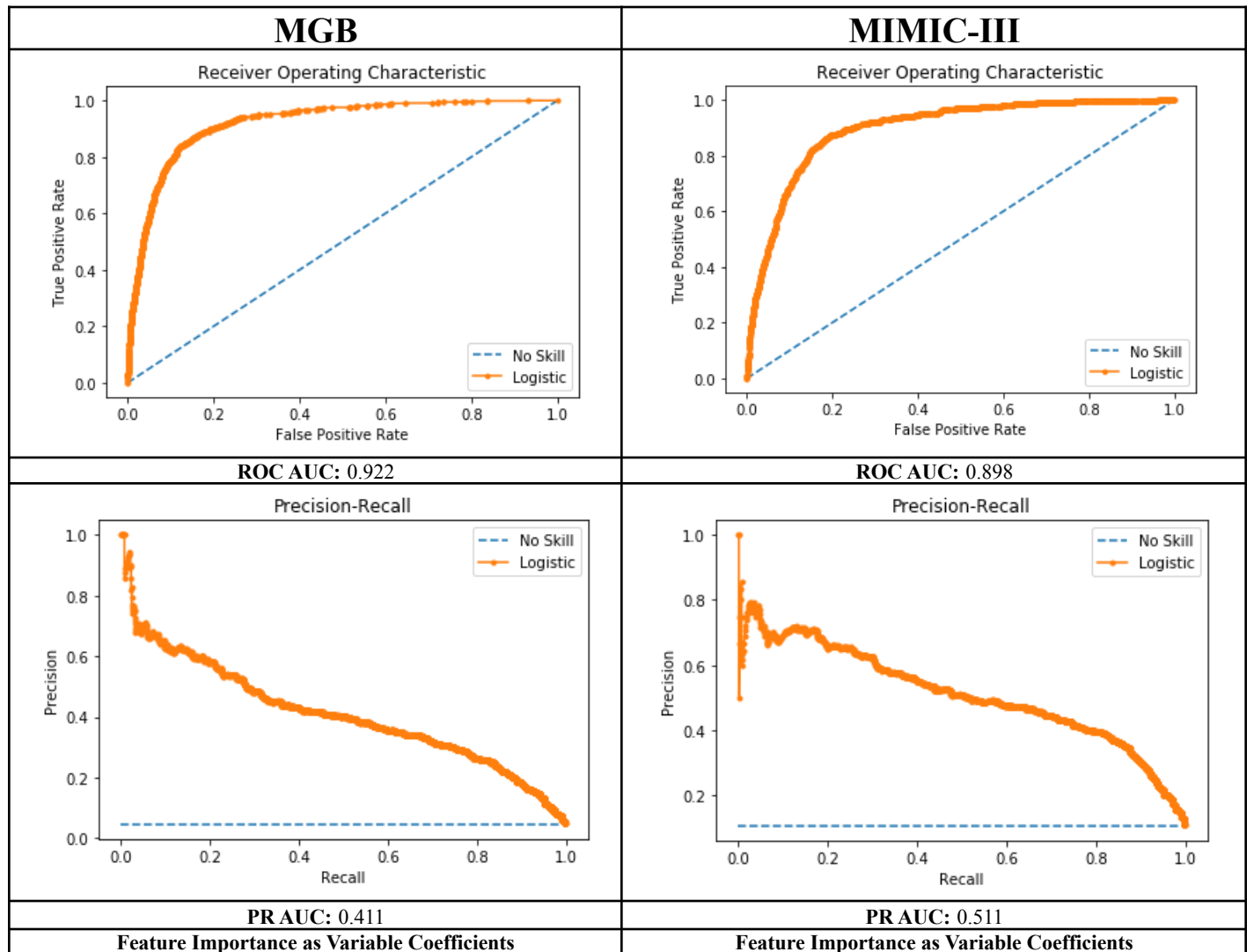
Best Logistic Regression Model Parameters		
	MGB	MIMIC-III
Score	0.957	0.909
C	0.01	0.01
Penalty	L2	L1
Solver	linlinear	saga

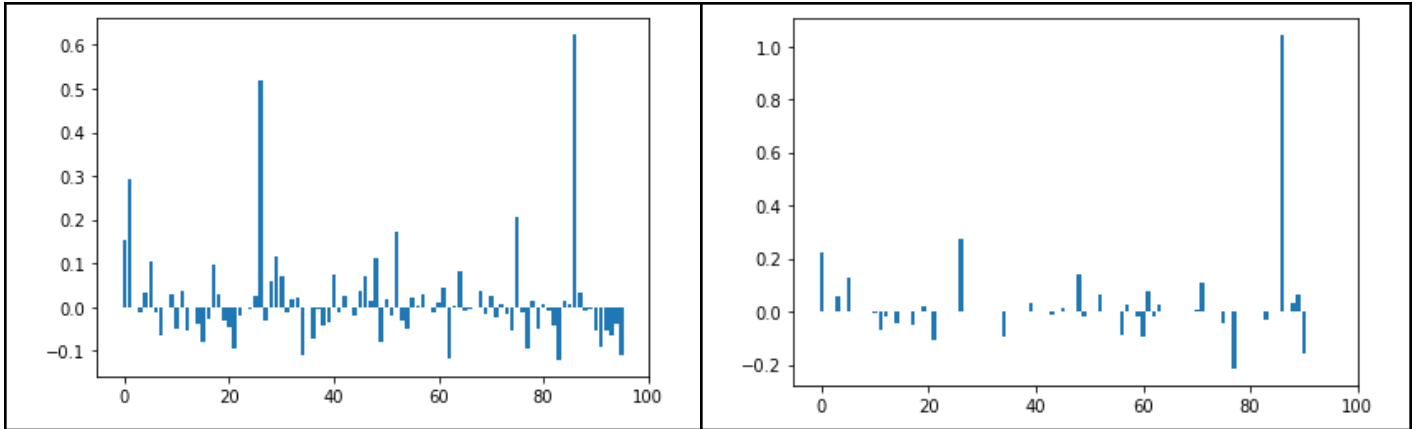
Logistic Regression Confusion Matrices				
	MGB		MIMIC-III	
	Predicted True	Predicted False	Predicted True	Predicted False
Actual True	113	675	182	605
Actual False	69	16,292	97	6,573

Best Logistic Regression Model Evaluation		
	MGB	MIMIC-III
Classification Error	0.04	0.09
Classification Accuracy	0.96	0.91
True Positive Rate, Recall	0.14	0.23
False Negative Rate	0.86	0.77
True Negative Rate	0.996	0.99
False Positive Rate	0.004	0.01
Precision	0.62	0.65
Negative predictive value	0.96	0.92
Balanced accuracy	0.57	0.61
F-1 score	0.23	0.34
Matthews correlation	0.28	0.35
ROC AUC	0.922	0.898
PR AUC	0.411	0.511

The logistic regression models perform better than trivial majority classifiers. The model trained on the MIMIC-III dataset has a better performance in balanced accuracy, F-1 score, and Matthews correlation coefficient (MCC). This is most likely due to the higher relative prevalence of the positive

class in the MIMIC-III dataset (10.29%) as compared to the MGB dataset (4.54%). This is a trend throughout the project in all models due to the higher level of imbalance of positives to negatives in the MGB dataset compared to MIMIC-III. This comparison is also evident in our ROC and Precision Recall curves below. The MCC will be most informative in comparing the multiple models due to the imbalanced characteristic of both datasets [28].





Above, the weights of the variables in the models above were compared and it was observed that the L1 regularization shrinks weights further than the L2 regularization, but that similar important features can be found with either method in either dataset. In Appendix Table D, it is observed that these similar important features related to the target variable were clinically meaningful and expected in both datasets, such as Renal Dysfunction and Chronic Kidney Disease.

## 4.2 Lasso Regression

Using a grid search to vary the constant that multiplies the L1 term,  $\alpha$ , the models with the best mean cross-validated score for each dataset were those with  $\alpha=0.01$ . Therefore, the best models were those that did not shrink most weights of features to zero. The model parameters selected from this grid search ( $\alpha=0.01$ ) were used to then test our resulting best Lasso regression model for prediction of a continuous eGFR value. Then the eGFR value is classified as class 1 or 0 for eGFR values  $<30$  and  $\geq 30$  respectively and compute the following classification evaluation metrics from the Confusion Matrices below.

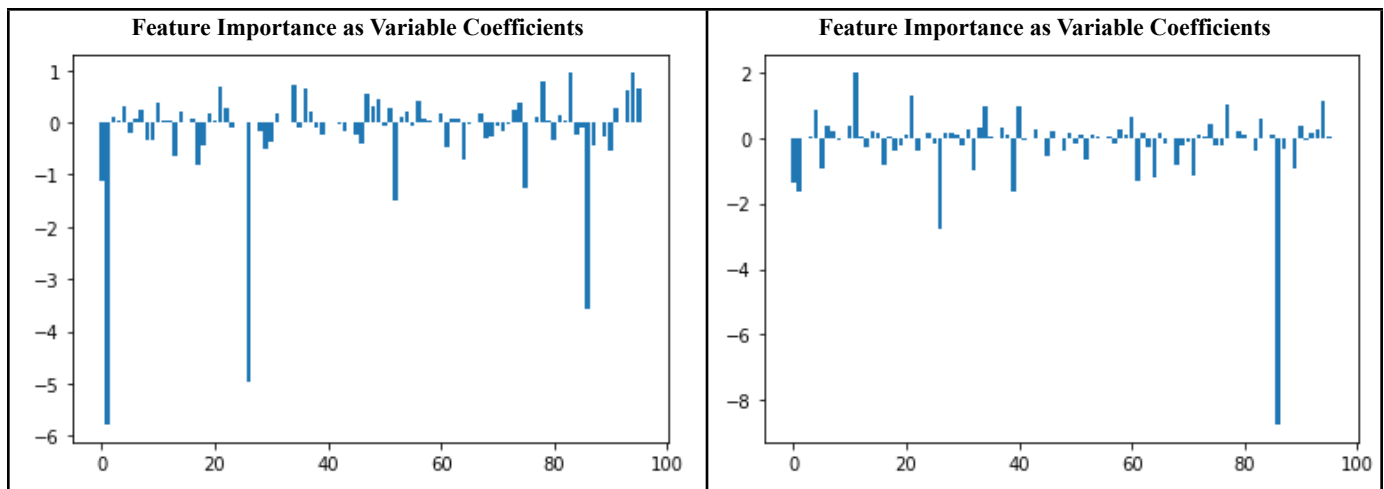
Best Lasso Regression Model Parameters		
	MGB	MIMIC-III
Score	0.957	0.909
$\alpha$	0.01	0.01

Lasso Regression Confusion Matrices				
	MGB		MIMIC-III	
	Predicted True	Predicted False	Predicted True	Predicted False
Actual True	61	727	150	640
Actual False	37	16,324	98	6,572

Best Lasso Regression Model Evaluation		
	MGB	MIMIC-III
Classification Error	0.04	0.10
Classification Accuracy	0.96	0.90
True Positive Rate, Recall	0.08	0.19

False Negative Rate	0.92	0.81
True Negative Rate	0.998	0.99
False Positive Rate	0.002	0.01
Precision	0.62	0.60
Negative predictive value	0.96	0.91
Balanced accuracy	0.54	0.59
F-1 score	0.14	0.29
Matthews correlation	0.21	0.30
MSE	241.109	264.643

The MSE suggests that the MBG Lasso regression model is a better estimator for eGFR than the MIMIC-III Lasso regression model, but when converting to classification, as observed in logistic regression, it is again observed that the MIMIC-III model performs better than the MGB model based on the balanced accuracy, F-1 score, and MCC. Below, it is observed differences in the assigned variable weights between these two Lasso regression models. It is important to note that there are some features that the MIMIC-III model seems to shrink or undervalue much more in comparison to the MGB dataset. This is most likely a characteristic due to the data itself since these two data sources differ substantially in their ability to capture each feature due to differences in care setting. The MGB dataset is enriched with a higher prevalence across all features and provides a more enhanced patient profile, but the MIMIC-III dataset is of a more severely sick and old population due to its patient care setting of critical care units of BIDMC. Whereas, the MGB data contains a more mixed population of both healthy and sick patients due to the variety of care settings (inpatient, outpatient, emergency room) that result from linking a large network of healthcare providers' electronic health records with insurance claims data. These differences in the patient populations and thereby data, are demonstrated in Table C of the appendix.



Additionally, here it is observed that Age, which is used in the calculation of eGFR, is very important in MGB, but not as relatively important in MIMIC-III. This is most likely due to imputation of the mean age

of 91.4 for patients older than 89, whose data was censored due to patient privacy regulations in the MIMIC-III dataset. Again, similar clinically meaningful predictors were found.

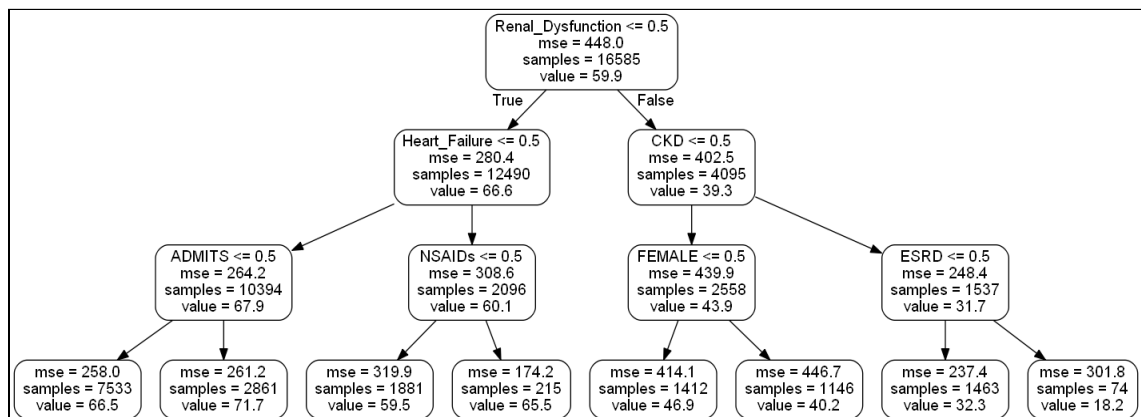
MGB Lasso Feature Importance	
Feature	Absolute Value of Variable Weight
Age	5.784
Chronic Kidney Disease	4.968
Renal Dysfunction	3.583
End Stage Renal Disease	1.486
Loop Diuretics	1.245
Female	1.123
Race - Black	0.949
Major Bleeding	0.941
Beta Blockers	0.807
All other variables	$\leq 0.774$

MIMIC-III Lasso Feature Importance	
Feature	Absolute Value of Variable Weight
Renal Dysfunction	8.739
Chronic Kidney Disease	2.754
Any hospitalization during prior 30 days	1.981
Age	1.652
Diabetes	1.652
Female	1.367
Heart Failure	1.330
Cancer	1.299
Hypertension	1.200
All other variables	$\leq 1.138$

### 4.3 Random Forest

The random forest regressor overall had fair results - the base model without hyper parameter had a MSE of 276.51 compared to an average baseline MSE of 451.1 where the baseline was taken as the average eGFR value across the data set. This was encouraging because it indicated that the base model had at least found some sort of indicator in the feature set which was better than random noise. One of the regression trees was pulled from the base model, but due to the size there was no optimal visualization. To generate a visualization, a mini-forest model was built, one of the trees from this mini-forest is shown below.

Example Regression Tree



The important features were extracted from the base model, which are based on the mean decrease in impurity for each feature, shown below [22]. This provides a useful ordering of how much influence each feature has on splitting the data set accurately. Note that Renal Dysfunction was the first node that the mini tree split on above as well as the most important feature in the table below, confirming that the tree seeks to split by highest selectivity first.

**Base Model Top Feature Importances**

<b>MGB</b>		<b>MIMIC-III</b>	
<b>Feature</b>	<b>Importance Score</b>	<b>Feature</b>	<b>Importance Score</b>
Chronic Kidney Disease	0.206	Renal Dysfunction	0.312
Age	0.139	Frailty Score	0.165
Frailty Score	0.137	Chronic Kidney Disease	0.022
Female	0.014	Heart Failure	0.018
Cancer	0.013	Female	0.017
Dipstick Urinalysis	0.013	Fluid Imbalance	.016
Glaucoma or Cataracts	0.013	H2 blockers	.015
Renal Dysfunction	0.013	EXPIRED	.014
All other variables	$\leq 0.012$	All other variables	$\leq 0.013$

From these tables, Renal Dysfunction was the most important parameter in the MIMIC dataset, while it was not a very important parameter in the MGB dataset. CKD was the most important parameter in the MGB data, but the third most important in the MIMIC data. These differences in importances are likely due to the fact that the data sources are fundamentally different, as shown in the confusion matrix at the end of this section - with significant imbalances for both positive and negative examples in the MIMIC data and MGB data respectively. One other important distinction in the data is that the MIMIC database had an average age of 91.10 vs the MGB database which had an average age of 73.6. This explains the fact that age is barely considered in the MIMIC regression forest since the effect of age is less noticeable when the entire dataset is older.

Once the base model was characterized, the model was improved via hyperparameter tuning. The process was identical for both the MGB and MIMIC data so the MIMIC process is detailed and results are listed for both. The first pass of tuning ran a 40 model split of parameter combinations randomly selected across the parameter set shown in the table below. Each model was trained using 5 fold cross validation, so a total of 200 models were trained for this step. Practically, this involves selecting a random value from each of the hyperparameters to generate a parameter set to pass to the model, and then running the cross validation on the model generated by those parameters. The models are then evaluated against each other by their minimum MSE to determine the best model from the parameter set.

Hyperparameter name	Values Considered
num trees	[100, 280, 460, 640, 820, 1000]
max depth	[10, 60, 110, 160, 210, None]
split samples	[2, 10, 20]
min leaf samples	[1, 2, 4]
bootstrap	[True, False]

This process resulted in a parameter set of {460,10,20,2,True} as the best set from the random search. These values were then taken as the basis for a grid search with some minor positive and negative adjustments. Note that the ‘bootstrap’ hyper parameter was removed because it was binary and the best parameter combinations contained only ‘true’. The grid search parameter values are shown below:

Hyperparameter name	Values Considered
num trees	[400, 460, 500, 540]
max depth	[8, 10, 12, 14]
split samples	[15, 20, 25]
min leaf samples	[1, 2, 3]

This required a search over 144 models running over 5 fold cross validation, for a total of 720 models trained. From this search, the best model on the MIMIC data ended up being {460,14,25,3}. On the MIMIC data, this perturb and grid search process was repeated once more to obtain the values in the following table (note that the second grid search MSE did not improve over the first grid search MSE):

Best Random Forest Regression Model Parameters		
	MGB	MIMIC-III
num trees	280	440
max depth	110	14
split samples	20	28
min leaf samples	4	3
bootstrap	True	True

From the final best models, the importance values were extracted below:

#### Best Model Top Feature Importances

MGB	
Feature	Importance Score
Chronic Kidney Disease	0.327
Age	0.182
Frailty Score	0.111
Renal Dysfunction	0.020
Female	0.013
ED Visit	0.013

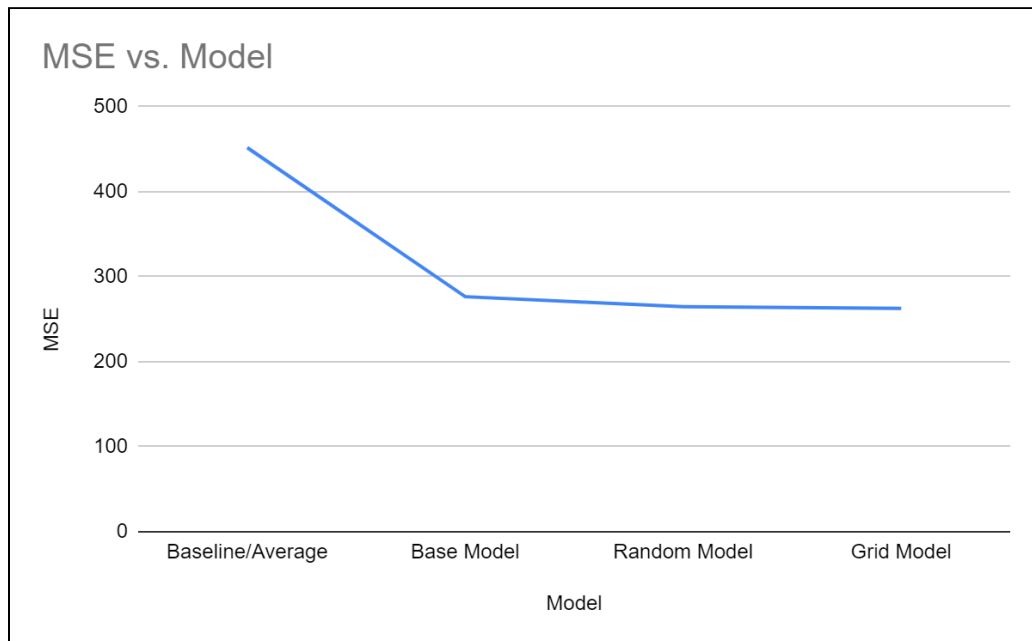
MIMIC-III	
Feature	Importance Score
Renal Dysfunction	.582
Frailty Score	.072
Chronic Kidney Disease	.04
Heart Failure	.029
Recent ER admission	.017
Female	.014

End Stage Renal Disease	0.011
Loop Diuretics	0.011
All other variables	$\leq 0.010$

Age	.012
NSAIDs	.011
All other variables	$\leq 0.01$

From these tables there are a few interesting observations. First, it can be seen that adjusting the hyper parameters boosted some importance scores: in MIMIC, Renal Dysfunction increased from 0.312 to 0.582; in MGB, CKD increased from 0.206 to 0.327. On the other hand, some features became less important to the model, such as frailty score in both the MIMIC and MGB data. This indicates that hyper parameter adjustment boosts the importance of the most selective features, and reduces the less important features to create a more predictive model.

Overall, the regression forest performed better than an average guess of the eGFR, and there was some minor improvement from the base model to the random hyperparameter adjusted model, but had minimal improvement from the random hyperparameter model to the grid search hyperparameter model, as shown in figure. This quick tapering in model performance over parameter adjustment combined with the quick drop in importance scores indicates that most of the features in our data are not very good indicators of eGFR, and there may be better features that weren't available in our dataset.



As a final analysis of the model the regression outputs are converted to classes and generated a confusion matrix as described above in Methods. These results are shown in the tables below. Overall the regression-turned-classifier regression forest model achieved a better than a trivial majority guess with a balanced accuracy of 0.54 and 0.61 on the MGB and MIMIC data respectively. This binary output model performed a bit better on the MIMIC data than on the MGB data. This is likely mostly due to the class



imbalance favoring negative examples in the MGB data set, and positive examples in the MIMIC data set. This resulted in an MGB model which had a very low chance of predicting true for any given input, resulting in an extremely low recall. Overall, the classification variant of the model on MIMIC had a better Matthew Correlation than the MGB model - which is a good indicator of performance regardless of class imbalance [28].

Random Forest Confusion Matrices				
	MGB		MIMIC-III	
	Predicted True	Predicted False	Predicted True	Predicted False
Actual True	86	1104	9901	155
Actual False	33	24500	867	267

Best Random Forest Regression Model Evaluation		
	MGB	MIMIC-III
Classification Error	0.04	0.09
Classification Accuracy	0.95	0.91
True Positive Rate, Recall	0.07	0.98
False Negative Rate	0.93	0.02
True Negative Rate	0.999	0.24
False Positive Rate	0.001	0.76
Precision	0.72	0.92
Negative predictive value	0.96	0.63
Balanced accuracy	0.54	0.61
F1-score	0.13	0.95
Matthews correlation	0.22	0.35
MSE	246.122	262.8

## 4.4 Neural Network

During the training process, as described in the Methods and Implementation section, a Randomized Grid Search was implemented over all of the available hyperparameters of the model. The table below summarizes the values that were used during the 15 randomized and sampled with replacement iterations.

Hyperparameter name	Description
num layers	[5, 7, 10, 12]
num neurons	[70, 100, 200, 300, 500]
activation	['relu', 'tanh']
dropout	[0.1, 0.2, 0.5]
optimizer	['Adam', 'RMSprop']
lr	[0.00001, 0.0001, 0.001, 0.01, 0.1]

Hyperparameters values considered for DNN model optimization

After the 15 iterations, it was determined that the best model consisted of 10 layers, 200 neurons per layer, hyperbolic tangent activation, 10% dropout, Adam optimizer, and 0.0001 learning rate. With the optimal parameters selected, the model was retrained on the training data, this time including the validation data. Once training was completed, the model was tested on the test set and the results are shown below.

Neural Network Confusion Matrices				
	MGB		MIMIC-III	
	Predicted True	Predicted False	Predicted True	Predicted False
Actual True	100	688	242	548
Actual False	70	16291	172	6498

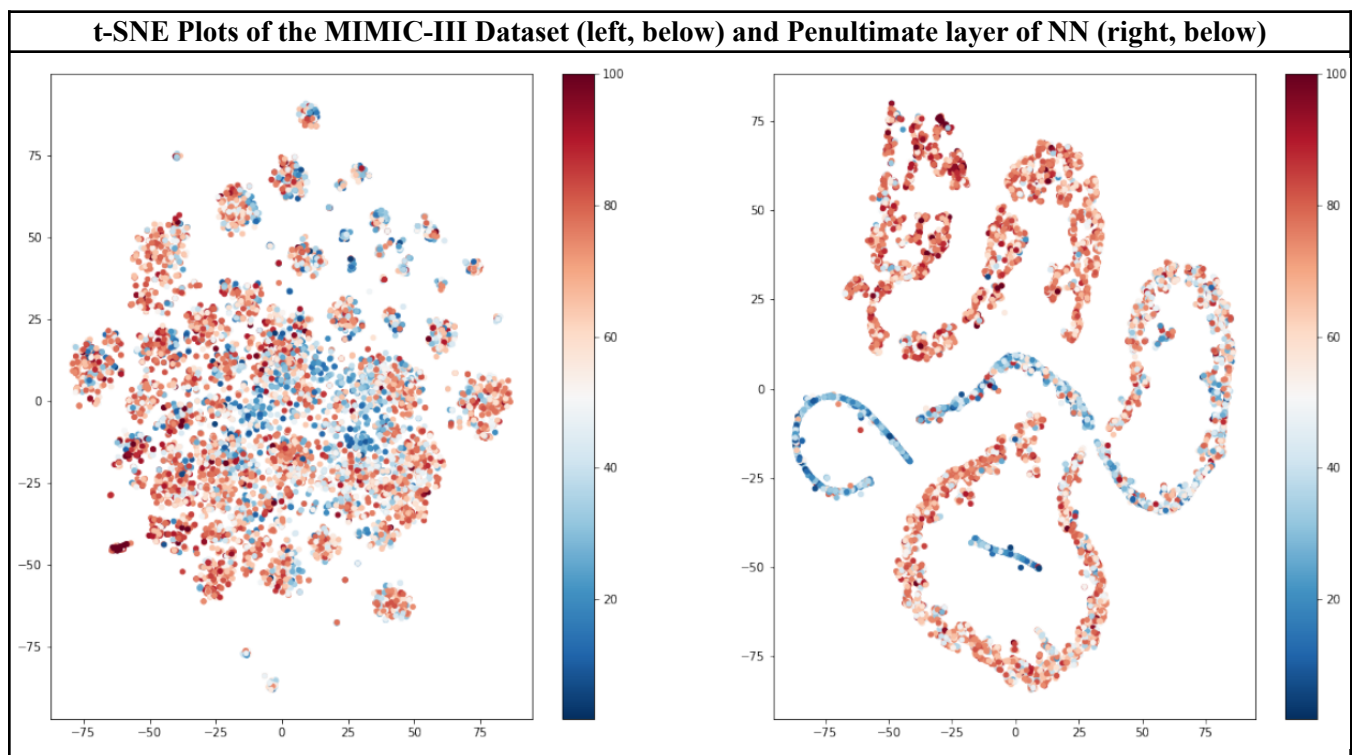
Best Neural Network Regression Model Evaluation		
	MGB	MIMIC-III
Classification Error	0.044	0.096
Classification Accuracy	0.956	0.903
True Positive Rate, Recall	0.127	0.306
False Negative Rate	0.873	0.694
True Negative Rate	0.995	0.974
False Positive Rate	0.004	0.024
Precision	0.588	0.585
Negative predictive value	0.959	0.922
Balanced accuracy	0.561	0.640
F1-score	0.209	0.402
Matthews correlation	0.259	0.377
MSE	246.31	270.11
MAE	12.74	12.91

As can be seen, the Neural Network performed well from an accuracy score perspective, however, when considering the imbalanced nature of the dataset, the results are more tenuous. In fact, it can be seen that the model tends to predict Negative class values ( $\text{eGFR} \geq 30$ ). This is perhaps not surprising as the vast majority of the dataset consists of values above 30 in both datasets. The results show the NN outperformed the simpler Random Forest and Linear Regression models by around 3% based on the balanced accuracy metric on the MIMIC-III with comparable results on the MGB data. However, the lack of a significant number of Positive class labels ( $\text{eGFR} < 30$ ) seemed to prevent the model from fully fitting to the data, and overfitting to the Negative class.

The Neural Network also performed significantly better on the MIMIC-III dataset than the MGB based on the balanced accuracy measurement. This can be explained by the further lack of eGFR values below 30 in the MGB dataset, preventing the model from properly learning the attributes of the Positive advanced CKD class.

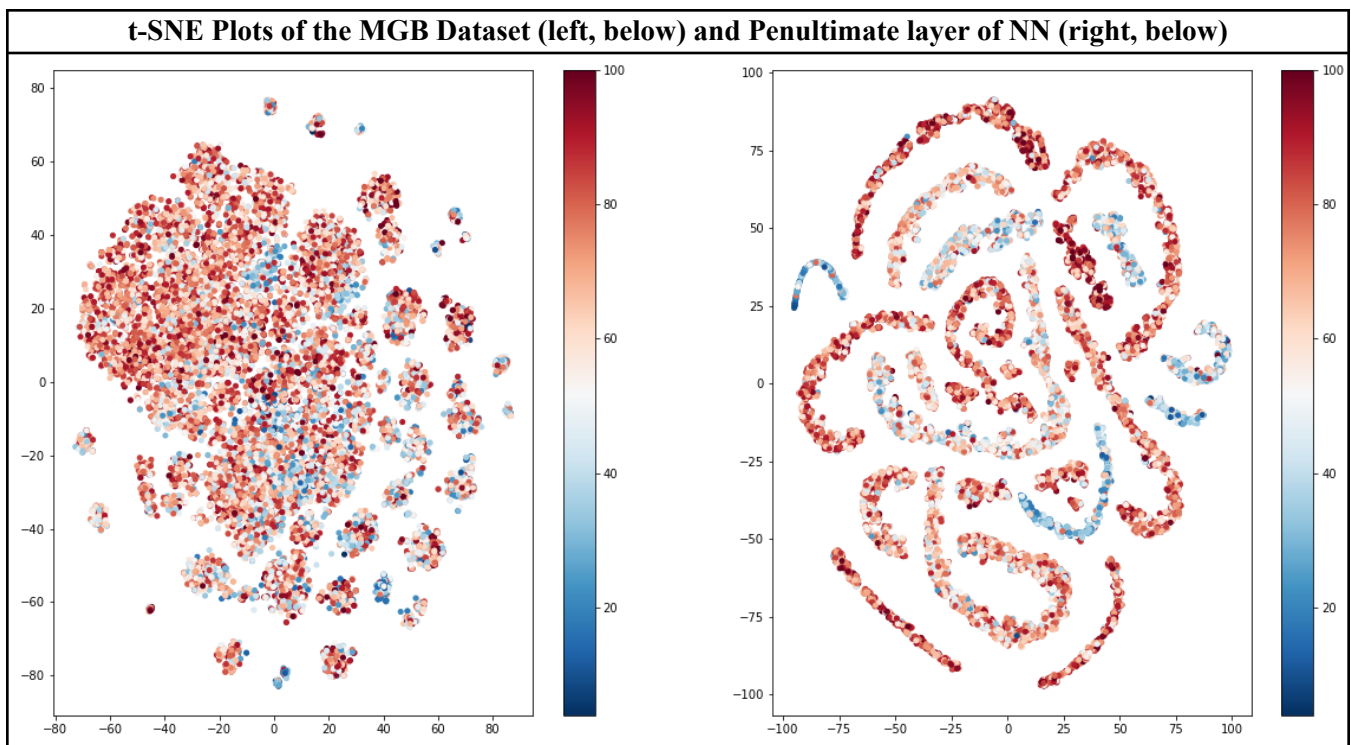
In the future, small to medium gains could possibly be realized by performing a more exhaustive grid search of hyperparameters if computing and time resources were not a concern, though it seems unlikely to vastly improve on the balanced accuracy numbers without more Positive class data points. In general, it is the authors' belief that a higher dropout rate (enforcing more regularization) coupled with a longer training cycle could help generalize the model better, particularly for heavily imbalanced datasets like the MGB.

Finally, in order to gain an understanding of how well the Neural Network was able to fit to the data, a dimensionality and clustering technique called t-SNE was used to convert predictions from the penultimate layer. t-Distributed Stochastic Neighbor Embedding (or t-SNE) converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. The result is that t-SNE will preserve local structures in the high dimensional data in the low level embedding, revealing potential underlying information about the nature of the dataset [26][27]. The t-SNE plot of the penultimate layer was then compared against the t-SNE of the raw data. The results of the dimensionality reduction can be seen below.



The t-SNE plot of the penultimate layer, seen on the right side of each plot, shows that the Neural Networks did in fact learn the underlying structure of the data. The high eGFR values (red colored) are all very close together and frequently in distinct clusters from the mid (white) and low (blue) eGFR values.

When compared with the dataset as a whole, the data is much more clustered around similar eGFR values. However, it is also noticeable that there aren't many of the low eGFR values and they frequently are located near mid eGFR values. Mid values are above our positive class threshold of 30 eGFR and thus this might give a visual explanation for why the model did not do as well from a balanced accuracy perspective.



## 4. Conclusions

Summary of Best Model Evaluation Metrics								
	Logistic Regression		Lasso Regression		Random Forest		Neural Network	
	MGB	MIMIC	MGB	MIMIC-	MGB	MIMIC	MGB	MIMIC
Classification Error	0.04	0.09	0.04	0.10	0.04	0.09	0.044	0.096
Classification Accuracy	0.96	0.91	0.96	0.90	0.95	0.91	0.956	0.903
True Positive Rate, Recall	0.14	0.23	0.08	0.19	0.07	0.98	0.127	0.306
False Negative Rate	0.86	0.77	0.92	0.81	0.93	0.02	0.873	0.694
True Negative Rate	0.996	0.99	0.998	0.99	0.999	0.24	0.995	0.974
False Positive Rate	0.004	0.01	0.002	0.01	0.001	0.76	0.004	0.024

Precision	0.62	0.65	0.62	0.60	0.72	0.92	0.588	0.585
Negative predictive value	0.96	0.92	0.96	0.91	0.96	0.63	0.959	0.922
Balanced accuracy	0.57	0.61	0.54	0.59	0.54	0.61	0.561	0.640
F-1 score	0.23	0.34	0.14	0.29	0.13	0.95	0.209	0.402
Matthews correlation	0.28	0.35	0.21	0.30	0.22	0.35	0.259	0.377
MSE	-	-	241.109	264.643	246.122	262.8	246.31	270.11

In general, all of the models were better estimators than a majority, trivial classifier. Due to the lack of positive advanced CKD data points, all models struggled to consistently predict True Positive cases. In the end, the Neural Network model was able to outperform both Random Forest and Logistic/Linear regression based on balanced accuracy. However, the Neural Network's training process was far more time and computationally expensive and thus provides potentially diminishing returns.

The dearth of low eGFR values was extremely restrictive when training the models to generalize well in terms of metrics like balanced accuracy. In the future, combining datasets (if possible based on usage restrictions) and only using a random sampling of data points with eGFR values above 30 may help increase the model's ability to correctly predict low eGFR. With additional time, the most important predictors could be selected with Linear regression (i.e. Lasso) or Random Forest first and then implement Logistic/Linear regression, Random Forest, and Neural Networks using that subspace of features to evaluate whether adding this additional step could improve classification balanced accuracy by removing features that provide no predictive value. While the lack of low eGFR values held back each of the models from predicting advanced CKD at a high rate, the regression outputs of the models provided relatively low Mean Absolute Error. Using the outputs from these regression models, without directly classifying advanced CKD, could provide physicians with additional data about the risk profile of their patient.

Another improvement which applies to all models universally is getting more and better data. Specifically, the MIMIC data set is highly biased since the only patients present in the data set are patients who were in critical care units. For this reason, models generated from the MIMIC data set are likely to only be valid on populations under similar conditions. This prevents generalization of the model and application of this model to a general population would likely have random or worse than random results. In order to remove this effect, a truly randomly sampled population data set would have to be acquired, or at least a data set which is representative of the general population.

Another possible improvement or experimentation that could be implemented in the scope of this project is studying the effect of varying the formula used to calculate eGFR. One potential formula to consider is MDRD [29]:

$$MDRD\ eGFR = 175 \times SCr^{-1.154} \times (Age)^{-0.203} \times 0.742[I = female] \times 1.212[I = black]$$

Additionally, both of these calculations of eGFR are reliant on race for their estimations. There is currently a heavy consideration to remove race from the eGFR calculations [30] since utilization of race in these calculations may produce a biased result and have concerning social, systemic consequences that lead to inaccuracy of measured eGFR in black patients and thereby affect care and health equity of these patients [31].

## **5. Individual Tasks (written by Nate Wilson)**

Lily worked quite extensively to obtain and clean the data. The process, especially for the MIMIC-III dataset, involved working with multiple CSV files, each representing different subsets of patient information. In addition, the patient information was published in unique hospital codes that she had to convert to usable patient features. Finally, she wrote the code for and tested the Linear and Logistic regression models implementing a grid search.

Lukas helped Lily with data cleaning of the MIMIC-III dataset. He wrote a few of the sub processes involved during the preprocessing including the calculation of the eGFR values. In addition, he wrote the code for the Random Forest models, implementing an exhaustive grid search methodology. He then tested and reported results from this model on the MIMIC-III dataset and passed on the code to Lily for her to run and report the results on the MGB dataset (due to restrictions on the use of the MGB data due to HIPAA).

I helped Lily implement grid search and cross validation on her Linear and Logistic regression models by creating a framework for her to use. I also wrote the code for the Neural Network, implementing a randomized grid search without the use of the Scikit learn library because of problems running it with Tensorflow. Finally, I tested and obtained the results for the Neural Network code on the MIMIC-III. Afterward, I passed on my code to Lily where she ran and reported the results on the MGB dataset (due to restrictions on the use of the MGB data due to HIPAA).

Finally, we all worked together frequently during our respective coding and project report writing time. We frequently had Zoom meetings where we could work and have the other group members available to answer any questions that may have come up.

## **Extension of 1R01LM013204-01A1**

The research conducted in this project will support the work of the Dr. Joshua K. Lin's R01 grant, 1R01LM013204-01A1, titled "Developing scalable algorithms to incorporate unstructured electronic health records for causal inference based on real-world data." The work conducted for this grant thus far has only been in the development of the algorithm to create the desired cohort of interest and evaluating

the accuracy of the dataset. No machine learning models have been applied within the scope of this grant thus far.

## References

- [1] J. Breiman, L. Random forests, *Machine Learning* 45(1), 5-32 (2001).
- [2] Bishop, Christopher M. *Pattern Recognition and Machine Learning*. New York :Springer, 2006.
- [3] Image of overfitting decision tree: <https://medium.com/@sametgirgin/Decision-tree-classification-in-9-steps-with-python-9cd5af04f4b8>
- [4] Tin Kam Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, Quebec, Canada, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.
- [5] Verikas, Antanas & Vaiciukynas, Evaldas & Gelzinis, Adas & Parker, James & Olsson, M. Charlotte. (2016). Electromyographic Patterns during Golf Swing: Activation Sequence Profiling and Prediction of Shot Effectiveness. *Sensors*. 16. 592. 10.3390/s16040592.
- [6] Rallapalli, Sreekanth & Suryakanthi, T.. (2016). Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm. 281-284. 10.1109/ICACCE.2016.8073762.
- [7] Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Acad Emerg Med*. 2016 Mar;23(3):269-78. doi: 10.1111/acem.12876. Epub 2016 Feb 13. PMID: 26679719; PMCID: PMC5884101.
- [8] Yang R, Plasek JM, Cummins M, Sward K. Predicting Falls among Community-Dwelling Older Adults: A Demonstration of Applied Machine Learning. *Computers, Informatics, Nursing*. 2020. (in press)
- [9] Geron, Aurelien. "Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow" O'Reilly (2019).
- [10] "What is a Perceptron: A Beginner's Tutorial for Perceptron" Simplilearn, <https://www.simplilearn.com/what-is-perceptron-tutorial>
- [11] Kim DH, Schneeweiss S, Lipsitz LA, Glynn R, Rockwood K, Avorn J. Measuring Frailty in Medicare Data: Development and Validation of a Claims-Based Frailty Index. *J Gerontol A Biol Sci Med Sci*. 2018; 73: 980-987. doi: 10.1093/gerona/glx229. PMID: 29244057; PMCID: PMC6001883 pmid: 29244057
- [12] Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 3, 160035 (2016). <https://doi.org/10.1038/sdata.2016.35>
- [13] "Estimated Glomerular Filtration Rate (eGFR)" by the National Kidney Foundation; <https://www.kidney.org/atoz/content/gfr>
- [14] Levey, A. S., & Stevens, L. A. (2010). Estimating GFR using the CKD Epidemiology Collaboration (CKD-EPI) creatinine equation: more accurate GFR estimates, lower CKD prevalence estimates, and better risk predictions. *American journal of kidney diseases : the official journal of the National Kidney Foundation*, 55(4), 622–627. <https://doi.org/10.1053/j.ajkd.2010.02.337>
- [15] Centers for Disease Control and Prevention. Chronic Kidney Disease in the United States, 2019. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2019.

- [16] Surya Krishnamurthy, Kapelesh KS, Erik Dovgan, Mitja Lustrek, Barbara Gradisek Piletic, Kathiravan Srinivasan, Yu Chuan Li, Anton Gradisek, Shabbir Syed Abdul. "Machine Learning Prediction Models for Chronic Kidney Disease using National Health Insurance Claim Data in Taiwan." medRxiv 2020.06.25.20139147; doi: <https://doi.org/10.1101/2020.06.25.20139147>
- [17] Yuan, Q., Zhang, H., Deng, T., Tang, S., Yuan, X., Tang, W., Xie, Y., Ge, H., Wang, X., Zhou, Q., & Xiao, X. (2020). Role of Artificial Intelligence in Kidney Disease. *International journal of medical sciences*, 17(7), 970–984. <https://doi.org/10.7150/ijms.42078>
- [18] E. Gregory Thompson, Adam Husney, Kathleen Romito, Tushar J. Vachharajani. "Acute Kidney Injury Versus Chronic Kidney Disease." *Healthwise*. 2019 August 11. <https://www.ummcc.org/health-library/aa106178#aa106178-sec>
- [19] Kueiyu Joshua Lin Daniel E. Singer Robert J. Glynn Shawn N. Murphy Joyce Lii Sebastian Schneeweiss. "Identifying Patients With High Data Completeness to Improve Validity of Comparative Effectiveness Research in Electronic Health Records Data." *Clinical Pharmacology & Therapeutics*. Volume 103 Number 5. 02 September 2017. <https://doi-org.ezproxy.neu.edu/10.1002/cpt.861>
- [20] Levey, Andrew S et al. "A new equation to estimate glomerular filtration rate." *Annals of internal medicine* vol. 150,9 (2009): 604-12. doi:10.7326/0003-4819-150-9-200905050-00006
- [21] Radivojac, Predrag,. *PERFORMANCE EVALUATION*. Nov. 2020, <https://www.ccs.neu.edu/home/radivojac/classes/2020fallcs6140/slidescs6140evaluation.pdf>. PowerPoint Presentation.
- [22] "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark" towards data science, <https://rb.gy/lmfip>
- [23] SciKit learn, sklearn.preprocessing.scale documentation, 2007 - 2020, scikit-learn developers (BSD License). <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.scale.html>
- [24] SciKit learn, Standardization, or mean removal and variance scaling documentation, 2007 - 2020, scikit-learn developers (BSD License). <https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-scaler>
- [25] sklearn.tree.DecisionTreeRegressor documentation, 2007 - 2020, scikit-learn developers (BSD License). <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>
- [26] "TSNE". Scikit Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>
- [27] L. van der Maaten. G. Hinton. "Visualizing Data using t-SNE". [https://lvdmaaten.github.io/publications/papers/JMLR\\_2008.pdf](https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf)
- [28] Boughorbel, S.B (2017). "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". *PLOS ONE*. 12 (6): e0177678. Bibcode:2017PLoSO..1277678B. doi:10.1371/journal.pone.0177678. PMC 5456046. PMID 28574989.
- [29] MDRD Study Equation. National Kidney Foundation. <https://www.kidney.org/content/mdrd-study-equation>
- [30] Kidney Disease, Race, and GFR Estimation. Andrew S. Levey, Silvia M. Titan, Neil R. Powe, Josef Coresh, Lesley A. Inker. *CJASN* Aug 2020, 15 (8) 1203-1212; DOI: 10.2215/CJN.12791019
- [31] Inzerro A. Flawed Racial Assumptions in eGFR Have Care Implications in CKD. <https://www.ajmc.com/view/flawed-racial-assumptions-in-egfr-have-care-implications>



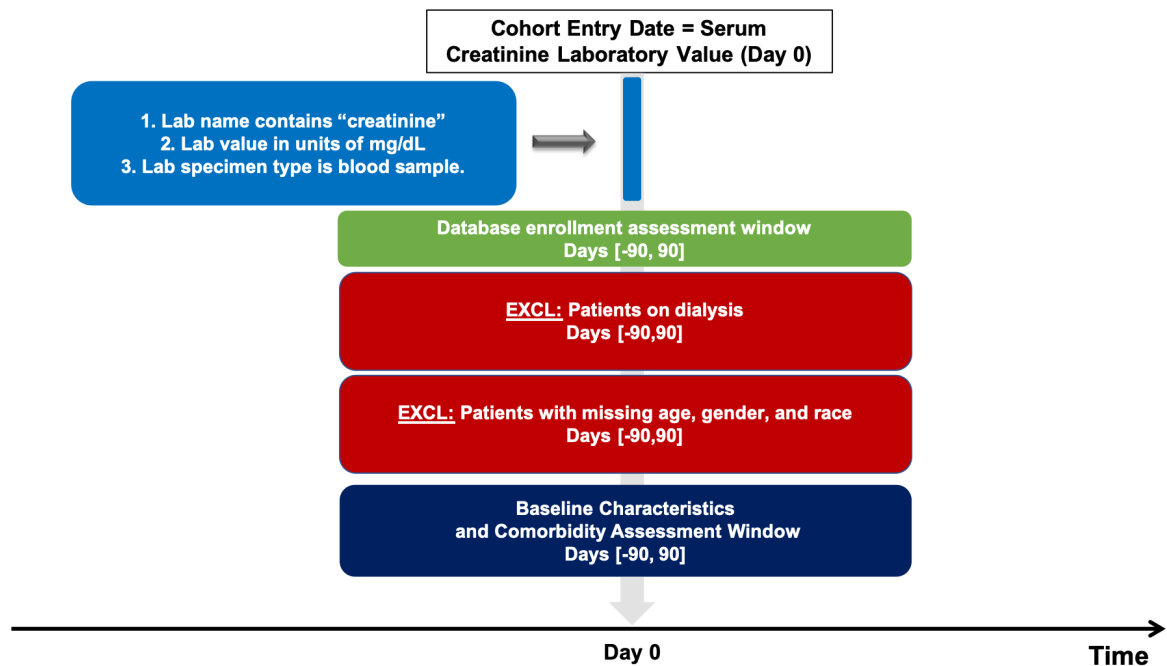


# Appendix

## Methods - Data

**Figure 1** - Study Design Diagram for Selection of Patients with recorded eGFR values and baseline characteristics

### Advanced Chronic Kidney Disease Phenotyping



**Table A.** Creatinine Cohort Entry Definition

ROW_ID	ITEMID	LABEL	FLUID	CATEGORY	LOINC_CODE
113	50912	Creatinine	Blood	Chemistry	2160-0

**Table B.** Dialysis Exclusion Criteria Definition

Code Type	Codes
ICD-9 Diagnosis	'V451', 'V4511', 'V4512', 'V560', 'V561', 'V568', 'V5681'
ICD-9 Procedure	'3995', '5498'
CPT-4 Procedure	'90945', '90947', '90999', 'S9339'

**Table C.** Patient Characteristics of MGB and MIMIC data measured 90 days before and 90 days after Serum Creatinine laboratory value measured

	<b>MGB</b>	<b>MIMIC</b>
<b>n</b>	85,743	37,297
Advanced Chronic Kidney Disease (eGFR < 30); n (%)	3896 (4.54)	3836 (10.29)
eGFR (mean (SD))	66.07 (19.78)	60.09 (22.19)
Female; n (%)	48961 (57.1)	16293 (43.7)
Age (mean (SD))	73.65 (7.85)	91.10 (4.32)
Race - Black ; n (%)	3890 (4.5)	2791 (7.5)
Frailty Score (mean (SD))	0.16 (0.06)	0.16 (0.04)
ACS/unstable angina; n (%)	4623 (5.4)	1637 (4.4)
ACE inhibitors; n (%)	6363 (7.4)	4215 (11.3)
Acute MI; n (%)	1740 (2.0)	3149 (8.4)
All-cause mortality; n (%)	4059 (4.7)	6092 (16.3)
Amputation; n (%)	528 (0.6)	227 (0.6)
Angina; n (%)	21355 (24.9)	979 (2.6)
Anti-arrhythmics; n (%)	2806 (3.3)	4625 (12.4)
Antiplatelet agents; n (%)	8896 (10.4)	0 (0.0)
Anxiety; n (%)	12601 (14.7)	1487 (4.0)
Any hospitalization during prior 30 days; n (%)	85215 (99.4)	10223 (27.4)
Any stroke; n (%)	13340 (15.6)	3627 (9.7)
ARBs; n (%)	12238 (14.3)	1950 (5.2)
Asthma; n (%)	8189 (9.6)	2412 (6.5)
Atherosclerosis; n (%)	26801 (31.3)	10586 (28.4)
Atrial Fibrillation; n (%)	17574 (20.5)	8901 (23.9)
Beta Blockers; n (%)	39298 (45.8)	21595 (57.9)
Bladder cancer, malignant and situ; n (%)	1874 (2.2)	99 (0.3)
Bladder Stones; n (%)	351 (0.4)	12 (0.0)
Bradycardia; n (%)	11504 (13.4)	2123 (5.7)
Cancer; n (%)	30745 (35.9)	5608 (15.0)
Cardiac Conduction Disorders; n (%)	5572 (6.5)	1111 (3.0)
Cardiovascular Stress Test; n (%)	671 (0.8)	0 (0.0)
Cellulitis or Abscess of Toe; n (%)	838 (1.0)	102 (0.3)
Cerebrovascular procedure; n (%)	612 (0.7)	247 (0.7)
Chronic Kidney Disease; n (%)	13127 (15.3)	3538 (9.5)
Excessive/frequent menstruation; n (%)	29 (0.0)	25 (0.1)
Hematuria; n (%)	5668 (6.6)	718 (1.9)
diagnosis indicating acute bleeding; n (%)	27552 (32.1)	7480 (20.1)
GI Bleed; n (%)	6088 (7.1)	2951 (7.9)
COPD; n (%)	14199 (16.6)	4164 (11.2)
Coronary Artery Disease; n (%)	28728 (33.5)	12544 (33.6)
Coronary atherosclerosis and other forms of chronic ischemic heart disease; n (%)	26128 (30.5)	10584 (28.4)
Coronary revascularization; n (%)	3383 (3.9)	6554 (17.6)
Cystoscopy; n (%)	3674 (4.3)	33 (0.1)
Cytology; n (%)	6803 (7.9)	0 (0.0)
Delirium; n (%)	7123 (8.3)	2722 (7.3)
Depression; n (%)	17928 (20.9)	3256 (8.7)

Diabetes; n (%)	23517 (27.4)	8206 (22.0)
Diabetes mellitus without mention of complications; n (%)	23052 (26.9)	6917 (18.5)
Diabetes with other ophthalmic manifestations; n (%)	2132 (2.5)	434 (1.2)
Diabetes with peripheral circulatory disorders; n (%)	3508 (4.1)	101 (0.3)
Diabetes with unspecified complication; n (%)	2127 (2.5)	73 (0.2)
Diabetic foot; n (%)	2569 (3.0)	592 (1.6)
Diabetic ketoacidosis; n (%)	187 (0.2)	478 (1.3)
Diabetic retinopathy; n (%)	2204 (2.6)	403 (1.1)
Dipstick urinalysis; n (%)	39409 (46.0)	0 (0.0)
Disorders of fluid electrolyte and acid-base balance; n (%)	20298 (23.7)	9993 (26.8)
Dorsopathies; n (%)	26372 (30.8)	1773 (4.8)
Drug abuse or dependence; n (%)	142 (0.2)	1693 (4.5)
Echocardiogram; n (%)	2890 (3.4)	2629 (7.0)
End Stage Renal Disease; n (%)	899 (1.0)	139 (0.4)
Falls; n (%)	5493 (6.4)	1081 (2.9)
Flu Vaccine; n (%)	22806 (26.6)	12 (0.0)
Foot Ulcer; n (%)	2679 (3.1)	601 (1.6)
Fractures; n (%)	8459 (9.9)	3433 (9.2)
Gangrene; n (%)	327 (0.4)	69 (0.2)
Gestational Diabetes; n (%)	1 (0.0)	7 (0.0)
Glaucoma or Cataracts; n (%)	25105 (29.3)	630 (1.7)
H2 blockers; n (%)	1393 (1.6)	15821 (42.4)
Heart Failure; n (%)	17347 (20.2)	8555 (22.9)
Hyperlipidemia; n (%)	53893 (62.9)	10697 (28.7)
Hyperosmolar hyperglycemic nonketotic syndrome; n (%)	591 (0.7)	60 (0.2)
Hypertension; n (%)	65464 (76.3)	19299 (51.7)
Hyperthyroidism; n (%)	1254 (1.5)	95 (0.3)
Hypoglycemia; n (%)	1097 (1.3)	106 (0.3)
Hypotension; n (%)	8205 (9.6)	1382 (3.7)
Ischemic Heart Disease; n (%)	28362 (33.1)	12170 (32.6)
Kidney and Renal Pelvis Cancer; n (%)	1226 (1.4)	170 (0.5)
Kidney Stones; n (%)	2982 (3.5)	155 (0.4)
Kidney Transplant; n (%)	385 (0.4)	251 (0.7)
Laser Photocoagulation and Vitrectomy; n (%)	218 (0.3)	17 (0.0)
Late effects of cerebrovascular disease; n (%)	5054 (5.9)	855 (2.3)
Liver Disease; n (%)	8945 (10.4)	3982 (10.7)
Loop Diuretic; n (%)	15272 (17.8)	17540 (47.0)
Lower Extremity Amputation; n (%)	528 (0.6)	328 (0.9)
NSAIDs; n (%)	10156 (11.8)	4148 (11.1)
ED Visit; n (%)	44161 (51.5)	0 (0.0)
Office Visit; n (%)	83268 (97.1)	454 (1.2)
Obesity; n (%)	13799 (16.1)	2071 (5.6)
Peptic Ulcer Disease; n (%)	1794 (2.1)	921 (2.5)
Valvular Disease; n (%)	8364 (9.8)	2003 (5.4)
Major Bleeding; n (%)	14400 (16.8)	7386 (19.8)
Peripheral Vascular Disease (PVD) or PVD Surgery; n (%)	11502 (13.4)	2308 (6.2)
PPI; n (%)	24446 (28.5)	20094 (53.9)
Renal Dysfunction; n (%)	20152 (23.5)	9232 (24.8)

Statins; n (%)	44241 (51.6)	13687 (36.7)
Type 2 Diabetes Mellitus; n (%)	23731 (27.7)	8206 (22.0)
Type 1 Diabetes Mellitus; n (%)	3978 (4.6)	775 (2.1)
ACE inhibitor; n (%)	28697 (33.5)	9767 (26.2)
DOAC; n (%)	1005 (1.2)	61 (0.2)
Warfarin; n (%)	12167 (14.2)	6035 (16.2)
Electrocardiogram; n (%)	58926 (68.7)	3 (0.0)

## Logistic Regression

**Table D.** Top Important Features as Described by the Absolute Value of Weight

<b>MGB</b>	
<b>Feature</b>	<b>Absolute Value of Variable Weight</b>
Renal Dysfunction	0.624
Chronic Kidney Disease	0.521
Age	0.292
Loop Diuretics	0.205
End Stage Renal Disease	0.171
Female	0.155
Major Bleeding	0.122
Hyperlipidemia	0.119
Diagnosis of Acute Bleeding	0.115
All other variables	≤0.111

<b>MIMIC-III</b>	
<b>Feature</b>	<b>Absolute Value of Variable Weight</b>
Renal Dysfunction	1.041
Chronic Kidney Disease	0.273
Female	0.225
NSAIDs	0.215
ACE inhibitors	0.160
Disorders of fluid electrolyte and acid-base balance	0.139
All-Cause Mortality	0.128
Kidney Transplant	0.108
Cancer	0.105
All other variables	≤0.098