Prediction of Acute Kidney Injury in Hospitalized COVID-19 Patients Using Machine Learning

By
David James Carbonello
B.A., Wheaton College (MA), 2019

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in the Department of Biostatistics at Brown University

PROVIDENCE, RHODE ISLAND

May 2021

This thesis by David Carbonello is accepted in its present form by the Department of Biostatistics as satisfying the thesis requirements for the degree of Master of Science

Advisor: Christopher Schmid Digitally signed by Christopher Schmid Date: 2021.04.01 10:54:05 -04'00'

Christopher H. Schmid, PhD

Reader: P Sarin

Pankaj Sarin, MD

Master's Program Director: Stavroula Chrysanthopoulou Digitally signed by Stavroula Chrysanthopoulou Date: 2021.04.01 11:18:45 -04'00'

Stavroula Chrysanthopoulou, PhD

Approved by the Graduate Council

Andrew G. Campbell, PhD, Dean of the Graduate School

Acknowledgement

I would like to first thank my advisor Dr. Christopher Schmid of the Department of Biostatistics at Brown University for providing consistent guidance and motivation throughout this research. Dr. Schmid's passion for statistics and dedication to his students is unmatched and I am extremely grateful for all the knowledge and insight he shared with me during our meetings to make this thesis possible.

I would also like to thank my thesis reader Dr. Pankaj Sarin of Brigham & Women's Hospital for dedicating his time and sharing his expertise to steer me in the right direction whenever I had questions. Dr. Sarin was instrumental in acquiring the data set and offered invaluable clinical perspective. I am also extremely thankful for his mentorship and continued support to prepare me for my future academic and professional endeavors.

Lastly, I would like to express my gratitude for my parents for their many sacrifices to allow me this opportunity and for their continuous encouragement throughout many years of study. This accomplishment would not have been possible without you.

David Carbonello

Table of Contents

List of Tables

List of Figures

1. Introduction

Originating in 2019, the highly infectious COVID-19 virus has become a global pandemic with 109M confirmed cases and 2.4M deaths worldwide [1]. The virus is caused by a newly identified coronavirus, SARS-CoV-2 (Severe Acute Respiratory Syndrome) and is known to have many physiological manifestations. The clinical spectrum of COVID-19 infection is broad, ranging from no symptoms to critical illness [2]. Initially thought of as a lung infection, the virus is now also understood to impact the cardiovascular system causing complications in the kidneys, brain, and other organs [3].

Recent studies show that acute kidney injury (AKI), a sudden episode of kidney failure that happens within a few hours or days, is associated with increased disease severity and mortality in hospitalized COVID-19 patients [4,5,18,19]. AKI causes a build-up of nitrogen waste products as measured by blood urea nitrogen (BUN) and serum creatinine levels (sCr), and impairs the kidney's ability to balance fluids in the body [6]. Unlike chronic kidney disease (CKD), where kidney function progressively declines over an extended period of time (>3months) and is irreversible, AKI is treatable when detected promptly [7]. The most common treatment of AKI is hemodialysis (commonly referred to as dialysis), where an artificial kidney (hemodialyzer) is used to remove toxins and excess fluids from the body while the kidneys heal [8,9]. The current diagnostic approach of AKI is based on increases in sCr and/or urine output over a specified time interval [2].

Latest data provides evidence of increased prevalence of AKI among hospitalized COVID-19 patients, with AKI occurring in >37.5% of fatal COVID-19 cases and in >50% of patients admitted to the ICU. Additionally, up to 20% of COVID-19 cases experiencing AKI require dialysis; this has led to shortages in dialysis supplies and medical equipment during spikes of the pandemic [5,10]. Therefore, identifying COVID-19 patients that are at higher risk of AKI is of great importance, and could help hospitals allocate resources more effectively and improve patient outcomes moving forward.

Prediction models of AKI occurrence in hospitalized COVID-19 patients have been developed and yielded high prediction accuracy, such as a study at Mount Sinai Hospital where XGboost was implemented and produced AUC of .79 [11]. However, few machine learning algorithms incorporate demographics, vitals, lab values, ICU admission, and pre-existing medical conditions in conjunction. With this in mind, the following study reports the implementation of logistic regression, random forest, and neural network models to predict AKI using data from electronic medical records of hospitalized COVID-19 patients at Brigham & Women's and Massachusetts General Hospitals in Boston, MA.

The aim of this study is to develop a tool to predict the risk of AKI among hospitalized COVID-19 patients within their first 48 hours of hospital admission using a variety of risk factors. A strong prediction tool will aid in the appropriate triage of individuals with high risk of AKI, and help hospitals be better prepared with medical equipment moving forward.

2. Methods

2.1 Study Population

The study population included all positive and presumptive positive COVID-19 cases admitted to Brigham and Women's Hospital and Massachusetts General Hospital in Boston, MA between January 1, 2020 and December 31, 2020. Electronic Medical Records (EMRs) were investigated retrospectively to identify occurrences of AKI. A positive COVID-19 case was defined using a positive reverse transcriptase polymerase chain reaction (RT-PCR) test. Presumptive positive cases were also included as positive COVID-19 cases in the study population, where a presumptive case was defined as a patient experiencing positive COVID-19 symptoms where results were pending or inconclusive. Patients who were younger than 18 years old or had pre-existing end stage kidney disease (Chronic Kidney Disease) were excluded. Additionally, patients with a length of hospital stay less than 24 hours were excluded. Only the first hospital encounter was included for patients who had more than one COVID-19 related hospital stay.

2.2 Data Collection and Preparation

All data on hospitalized COVID-19 patients from January 1, 2020 to December 31, 2020 were queried from the Enterprise Data Warehouse (EDW), a database containing information extracted from the Epic Electronic Medical Record (EMR) of Brigham & Women's Hospital (BWH) and Massachusetts General Hospital (MGH) in Boston, MA. Specific information on demographics, vitals, lab values, ICU admission, and pre-existing conditions were extracted from the EDW.

Demographics included patient's age and sex. Vitals included systolic blood pressure

(SBP), diastolic blood pressure (DBP), body temperature, heart rate, body mass index (BMI), and respiration rate.

Lab values included as predictors were potassium, calcium, albumin, bicarbonate, chloride (CHL), white blood cell count (WBC), glucose, absolute neutrophil count (ANEU), blood test for c-reactive protein (CRPT), lactate dehydrogenase (LDH), international normalized ratio (INR), and prothrombin time (PT). Only lab values recorded 24 hours prior to admission and within the first 48 hours after admission were used in the analysis. The average lab values during this interval were calculated for patients and used as predictors. International normalized ratio, prothrombin time, and c-reactive protein were included as binary variables based on whether or not the patient received the test of interest. Tests were dichotomized rather than including their results because only a small proportion of patients had the tests performed, and receiving the test was likely indicative of more severe cases. All lab values of serum creatinine were used to identify the occurrence of AKI at any point during a patient's hospitalization based on an increase of 0.3mg/dL or more within 48 hours or by 1.5 times baseline or more within the last 7 days.

ICU on admission status was defined as a binary variable based on whether or not a patient was admitted to the ICU either directly or within the first 48 hours of hospital admission

Pre-existing conditions for patients were defined using the latest International Classification Disease Codes (ICD-10 codes). Comorbidities and their corresponding ICD-10 code included as predictors were Hypertension (I10), Congestive Heart Failure (I50.9), Liver Disease (K76.9) and Peripheral Vascular Disease (I73.9). A complete list of variables and their respective definitions are shown in Table 1.

Data validation was performed for extreme values of BMI, body temperature, white blood cell count, glucose, lactate dehydrogenase, heart rate, and respiration rate by confirming recorded values in EDW were consistent with Epic EMR. Outliers were identified based on comparison to reference ranges shown in Table A1 and visual inspection of continuous distributions. Outliers produced from probable data entry errors resulting from improper units were adjusted accordingly. Extreme lab values and vital recordings that were called back in Epic reports were considered as missing data. Called back lab values are usually extreme values that need immediate action, or specimens that are insufficient for a sample to be run properly. For patients with missing records of BMI on the hospital encounter of interest, the most recent historical BMI recorded for the patient was used.

Table 1: Description of Predictors

Variable Name	Type	Description
Potassium	continuous	avg Potassium lab value in mmol/L from 24hours prior-48 hours after hospital admission
Calcium	continuous	avg Calcium lab value in mg/dL from 24 hours prior-48 hours after hospital admission
Chloride	continuous	avg Chloride lab value in mmol/L from 24 hours prior-48 hours after hospital admission
Bicarbonate	continuous	avg Bicarbonate lab value in mmol/L from 24 hours prior-48 hours after hospital admission
Albumin	continuous	avg Albumin lab value in g/dl from 24 hours prior-48 hours after hospital admission
WBC	continuous	avg White Blood Cell Count in k/uL lab value from 24 hours prior-48 hours after hospital admission
Glucose	continuous	avg Glucose lab value in mg/dL from 24 hours prior-48 hours after hospital admission
LDH	continuous	avg Lactate Dehydrogenase lab value in U/L from 24 hours prior-48 hours after hospital admission
ANEU	continuous	avg Absolute Neutrophil Count lab value in k/uL from 24 hours prior-48 hours after hospital admission
Age	continuous	Age in years of patients on admission (defined as difference in hospital admit date and date of birth)
SBP	continuous	Systolic blood pressure recorded in mmHg recorded on admission
DBP	continuous	Diastolic blood pressure recorded in mmHg recorded on admission
Body Temperature	continuous	Body temperature recorded on admission
Heart Rate	continuous	Heart rate in beats per minute recorded on admission
BMI	continuous	Body Mass Index kg/m^2 recorded on admission (or most recent historical recording available)
Respiration Rate	continuous	Respiration rate in breaths per minute recorded on admission
INR	binary	International Normalized Ratio, 1= had test, 0=did not test
PT	binary	Prothrombin time, 1= had test, 0= did not have test
CRPT	binary	Blood test for C-reactive protein, 1=had test, 0=did not have test
Male	binary	1=male, 0=female
Hypertension	binary	1=having pre-existing Hypertension, 0= not having pre-existing
Diabetes	binary	1=having pre-existing Diabetes, 0= not having pre-existing Diabetes
Liver Disease	binary	1=having pre-existing Liver Disease, 0= not having pre-existing Liver Disease
Peripheral Vascular Disease	binary	1=having pre-existing Peripheral Vascular Disease, 0= not having pre-existing Peripheral Vascular Disease
ICU	binary	1= patient admitted to ICU, 0=patient not admitted to ICU
AKI	binary	Outcome variable of AKI 1=having AKI, 0=not having AKI. Defined using KDIGO criteria

2.3 Definitions of Outcomes:

The outcome variable of Acute Kidney Injury was defined using the following criteria from the Kidney Disease Improving Global Outcomes (KDIGO) guidelines:

1.) Increase in serum creatinine by 0.3mg/dL or more within 48 hours **or**

2.) Increase in serum creatinine by 1.5 times baseline or more within last 7 days

Baseline serum creatinine was defined as the first reading of serum creatinine within the last 7 days. All available recordings of serum creatinine following lab admission were included, and therefore all possible 7-day intervals were considered when defining AKI from criteria two.

2.4 Statistical Analysis

Methods considered for developing a classifier to predict AKI occurrence in hospitalized COVID-19 patients included logistic regression, random forest, and artificial neural network. Logistic Regression was chosen for its probabilistic interpretation, popularity in COVID-19 research, and competitive performance when compared to other machine learning algorithms used in classification tasks [22]. Random Forest and Neural Networks were chosen as alternative modeling approaches to allow more flexible relationships between prediction features and outcome, and to see if a higher prediction accuracy could be attained. Receiver operating characteristic curves and the area under the curve were plotted for the corresponding models to observe true positive rates and false positive rates across all possible probability thresholds. The Youden index was chosen as the optimal probability threshold to classify AKI status in the logistic regression, random forest, and neural network to provide the best tradeoff between sensitivity and specificity [11]. All

models were fit using a response variable of AKI, which was a binary variable corresponding to disease status. To increase generalizability of the study, data were randomly split 75% into a training set (3,817 observations) for model development, and the remaining 25% (1,273 observations) into a test set to assess model performance. To account for missing data, multivariate imputation by chained equations (MICE) was performed on both the training and test sets to produce 5 imputed datasets respectively.

2.4.1 Logistic Regression

Two separate logistic regression models were developed with and without 2-way interactions. The model without interactions was developed using a forward stepwise regression algorithm to perform variable selection when fitting the model on the 5 imputed training sets. All 26 predictors were included in the variable selection process for the main effects model. For each set of training data, the optimal model size was determined using 10-fold cross-validation with log-likelihood as the evaluation criterion. Resulting regression coefficients and standard error estimates were pooled using Rubin's Rule for the five sets of imputed training data to produce the main effects model. All pairwise interactions between features included in the main effects model were then investigated and selected via cross-validation. The forward selection algorithms for main effects and interactions via cross-validation are shown below:

Forward Stepwise Selection Algorithm for Main Effects via Cross-Validation

I. Divide the data randomly into K equally sized sets. For k in 1,…,K
    a. Treat set k as the holdout test set and the other K-1 sets as the training data

b. Construct the null model, M_{k0}, which contains no predictors and compute model evaluation index $E_{k,0}$ (log-likelihood)

c. Assume number of predictors is p. For j = 0, …, p-1:

i. Consider all p − j models that augment the *j* predictors in M_{kj} with one additional predictor; compute model evaluation index for each of p-j models fit to the training data

ii. Choose the best among these p − j models, and call it $M_{k,j+1}$

iii. Apply model $M_{k,j+1}$ on the holdout test set, and compute model evaluation index $E_{k,j+1}$

II. Compute average test performance $\bar{E}_j$ for j = 0, 1, …, p across K test sets as $\bar{E}_j = \frac{1}{K}\sum_{k=1}^{K} E_{kj}$ and

choose m=$\underset{j}{\operatorname{argmin}}\, \bar{E}_j$

III. Choose the final model as the result of forward stepwise regression on the whole data set stopping when the model has *m* variables.

Forward Stepwise Selection Algorithm for Interactions via Cross-Validation

Interaction terms were selected by applying the forward stepwise procedure below for all possible pairwise interaction

terms after running the main effects algorithm above.

I. Use the same division of k data sets. For k in 1,…,K

a. Treat set k as the holdout test set and the other K-1 sets as the training data

b. Take model $M_{k,j}$, *j=2,…,p* obtained from the previous algorithm as the null model with no interactions and define $M_{k,j,0} = M_{k,j}$. For each j, $q_j = \binom{j}{2}$ potential pairwise interactions are considered

c. For i = 1, …, q_j

i. Consider all q_j – i+1 models that augment the j + i-1 predictors (j main effects + i-1 interactions) in $M_{k,j,,i-1}$ with one additional interaction pair; compute the model evaluation index for each of q_j – i+1 models

ii. Choose the best model among these q_j – i+1 models, and call it $M_{k,j,i}$

iii. Apply model $M_{k,j,I}$ on the holdout test set, and compute model evaluation index $E_{k,j,i}$

II. Compute average test performance $\bar{E}_{ji}$ across K test sets as $\bar{E}_{ji} = \frac{1}{K}\sum_{k=1}^{K} E_{kji}$ for each j= 2,…,p and i= 1,…,q_j, and choose (m,n) = $\underset{j}{\operatorname{argmin}}\, \bar{E}_{ji}$ as optimal model size where *m* is the number of main effects and *n* is the number of interactions in the model

III. Choose final model as result of forward stepwise regression on whole data set stopping when the model has *m* variables and then *n* interactions among these *j* variables.

2.4.2 Random Forest

A random forest classifier was implemented as an extension of a single classification tree for its proven ability to make significant improvements to traditional CART methods. The algorithm uses a bootstrap aggregation (bagging) procedure and a random subset of features to fit multiple decision trees, resulting in decreased variance and decorrelation of trees. Classification predictions from random forests are determined using a majority vote rule in terminal lead nodes [13]. The random forest classifier was trained on the 5 imputed training sets using 10-fold cross-validation for hyperparameter tuning, with selection based on the combination of hyperparameters that yielded the highest average area under the ROC curve produced across each fold. Hyperparameters tuned during cross-validation included the number of trees (10,25,50,100), number of features considered at each split (square root, log base 2 of the total number of features), and fit criterion (Gini statistic, entropy). The model structure given by the resulting best combination of hyperparameters determined by cross-validation was used to make predictions on the holdout test sets.

2.4.3 Neural Network

A feed forward neural network was developed with the architecture of a multilayer perceptron neural network (MPNN), containing one input and output layer, and one or more hidden layers where each node was fully connected. This structure can lead to more

effective decision making by detecting and using hidden information that may be present in high volumes of data. An example of a MPNN structure with two hidden layers is shown in Figure 1. The blue circles serve as the input layer and correspond to the predictors included in the model. The gray circles are the neurons that form the hidden layers and connect successive layers. The red node is the output layer corresponding to the predicted response [14].

Figure 1. Multilayer Neural Network Structure

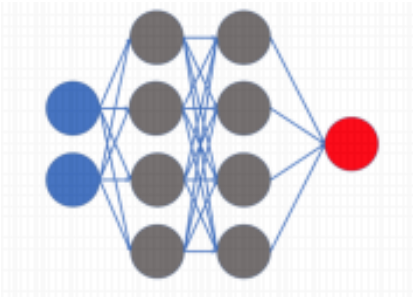
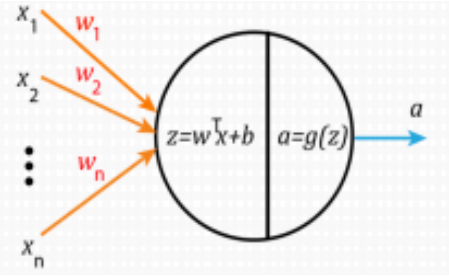


Figure 2: Mathematical Representation of 1 Neuron



The mathematical structure of an individual neuron is shown in Figure 2 where $x_1 \ldots x_n$ are a specific vector of inputs that are multiplied by weights $w_1 \ldots w_n$. A bias, b is then added to the weighted sum of inputs as shown in Equation 1:

$$z = \sum_{i=i}^{n} x_i w_i + b \qquad (1)$$

Here z is the net input fed through a specific node that is passed through an activation function $g(z)$ as shown in Equation 2:

$$a = g(z) = g(\sum_{i=1}^{n} x_i w_i + b) \qquad (2)$$

where a is the output that is passed on to other neurons [16]. The default Rectified Linear

Unit activation function (ReLU) was used for the input layer and hidden layers, and the sigmoid activation function used in the final hidden layer. The ReLU activation function was chosen for its computational efficiency and widespread use for deep learning tasks. The ReLU activation function is shown in Figure 4. The sigmoid activation function was used to restrict the outcome between 0 and 1 for a probabilistic interpretation and is shown in Figure **3.** The structure of the neural network was determined using 10-fold cross-validation for hyperparameter tuning using area under the ROC curve area as the evaluation criteria. The hyperparameters considered during the cross-validation process included the number of hidden layers (1,2), nodes within each hidden layer (10, 50, 100, 200), number of epochs (30,40), and alpha (.001,.0001). Alpha is the learning rate and controls how much model responds to error estimates while updating the weights during backwards propagation. Epochs are the number of times which the algorithm is passed through the training data [24]. The Adam optimizer algorithm was used as an alternative to stochastic gradient descent to minimize the binary loss function and recursively update weights when fitting the neural network [17]. All input features in the neural network were standardized to improve model performance and efficiency.

Figure 3. Sigmoid Activation Function

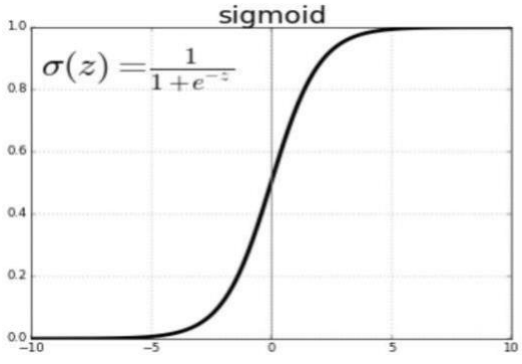
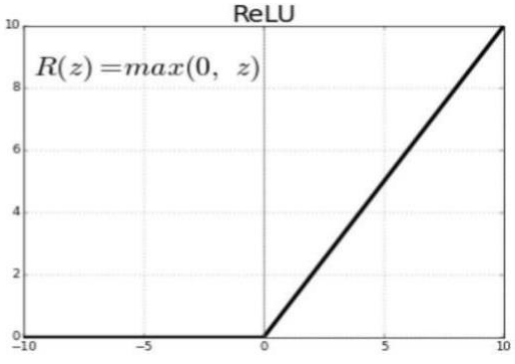


Figure 4. Relu Activation Function



3.1.1 Model Evaluation

Prediction metrics including accuracy, misclassification rate, sensitivity, specificity, area under the ROC curve (AUROC), and area under the precision recall curve (AUPRC) were aggregated across the 5 imputed test sets using Rubin's Rule and reported for the respective models. Accuracy measures the proportion of individuals that were correctly labeled into the respective classes and misclassification rate measures the proportion of patients that were incorrectly classified. Sensitivity, or true positive rate, measures the proportion of patients with AKI that are correctly classified as having AKI. Specificity, or true negative rate, measures the proportion of patients without AKI who are correctly classified as not having AKI. AUROC measures the performance of a classifier by comparing the true positive rates and false positive rates across all possible probability thresholds, where a larger AUROC corresponds to a stronger classifier. AUPRC measures the performance of a classifier comparing the sensitivity (recall) and positive predictive value (precision) across all possible classification thresholds [20]. This metric is often used in addition to AUROC and is useful for unbalanced data sets where the number of positives is much greater than the number of negatives. The Hosmer-Lemeshow Test was used to assess the goodness of fit for the respective models where individuals are binned into different subgroups based on predicted probabilities. A Chi-squared statistic is then calculated to compare the observed event rates and average predicted event rates in each bin, where a large statistic (with lower p-value) indicates a poorer model fit [23].

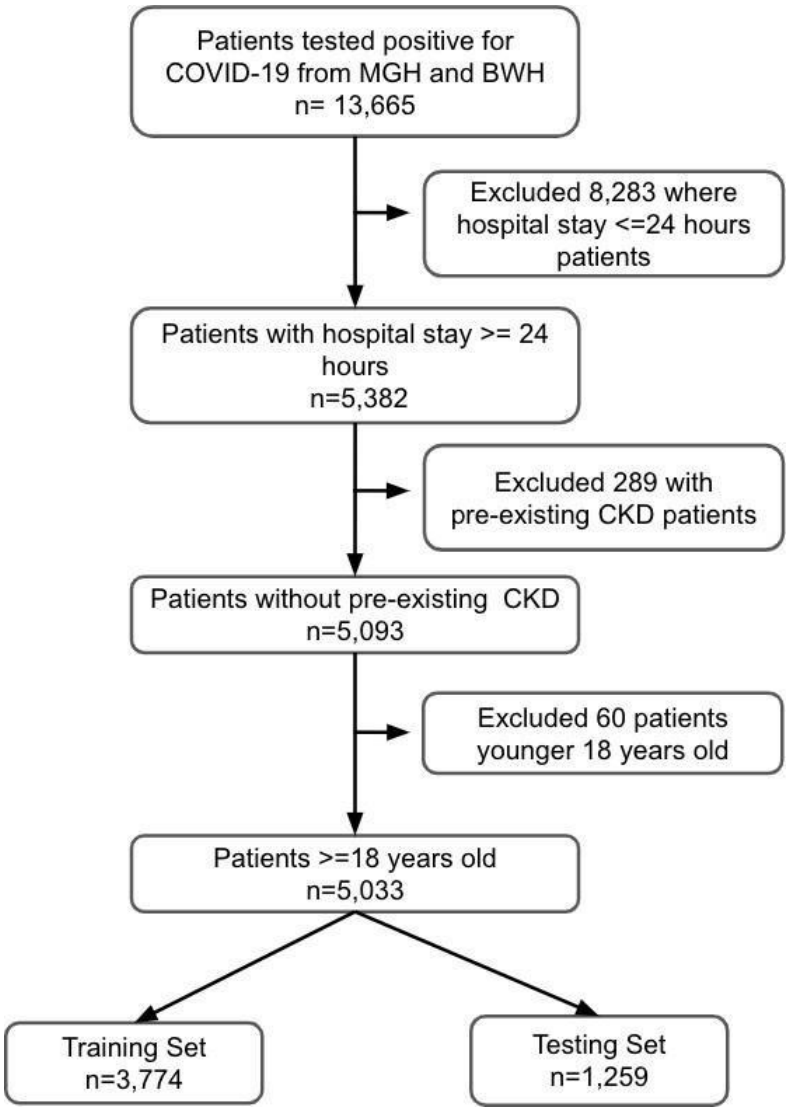
3.1.2 Statistical Software

The two separate logistic regression models with and without 2-way interactions were developed using R3.6.1. Both the neural network and random forest were fit using Python 3.7.6. The GridSearchCV function from the scikit-learn library in Python was used to perform 10-fold cross-validation for hyperparameter tuning for both the neural network and random forest. The mice package in R3.6.1 was used to perform multiple imputation by chained equation on missing values.

3. Results

The study population consisted of 5,033 confirmed or presumptive positive COVID-19 patients after removing patients based on exclusion criteria as shown in Figure 5. In total, 8,283 patients were excluded for not having a hospital stay of at least 24 hours, 289 patients were excluded for having pre-existing CKD, and 60 patients younger than 18 years old were excluded.

Figure 5. Flow Diagram



Standard summary statistics of continuous predictors were calculated and tabulated in Table A1. Histograms of continuous variables show distributions and potential outliers in data and are shown in Figure A1. The majority of continuous variables were approximately normally distributed. Variables exhibiting heavy positive skew included glucose, lactate dehydrogenase, white blood cell count, and absolute neutrophil count. Natural logarithmic transformations were imposed on these variables to normalize their respective distributions

and improve model performance (Figure A2). Variables that contained outliers based on histograms and comparison to normal reference ranges were body temperature, respiration rate, white blood cell count, glucose, and lactate dehydrogenase. The proportion of missing data within each variable before imputation is shown in Figure A3. The data were assumed to be missing at random so that missing values were not related to any unobserved factors. Figure A4 shows the comparisons of variable distributions when specific variables were and were not missing. The similarity between boxplots suggests evidence that there is no relationship between missing and observed data. Potassium, calcium, chloride, bicarbonate are measured from the same blood test, and therefore are all expected all to be missing if the patient did not receive the test as observed. The incidence of AKI within specific subgroups is shown in Table A2. The overall incidence of AKI in the observed study population was 24.7%. More AKI was observed for patients with pre-existing hypertension (30.6%) compared to those without pre-existing hypertension (20.6%). Additionally, 59.64% of patients admitted to the ICU on admission experienced AKI as compared to 19.11% of patients experiencing AKI that were not admitted to the ICU on admission.

The resulting best parameter combination determined by cross-validation in the random forest was using the Gini statistic criterion, 125 trees, and considering the square root of the total number of predictors at each split. The optimal neural network determined by cross validation included 2 hidden layers, where the number of nodes in each layer respectively were 100 and 25. The input layer contained 25 nodes corresponding to each of the predictors included in the model. The optimal learning rate and number of epochs were .001 and 40 respectively. The optimal size of the main effects logistic regression model determined by cross-validation contained 18 predictors. The only significant pairwise interactions between these main effects were between systolic and diastolic blood pressure.

Aggregated metrics of prediction accuracy, sensitivity, specificity, AUROC, and AUPRC, across the 5 imputed test sets are shown for the respective models in Table 2. The summary output of the main effects logistic regression model determined via cross-validation that was used for model evaluation and its summary output is shown in Table 3.

Table 2. Prediction Metrics for LR, RF and NN

Model	Accuracy	MissClass	Sensitivity	Specificity	AUROC	AUPRC
Logistic Regression	0.766	0.234	0.642	0.804	0.781	0.532
Random Forest	0.759	0.241	0.653	0.791	0.795	0.564
Neural Network	0.772	0.228	0.643	0.811	0.786	0.561

Table 3. Main Effects Stepwise Logistic Regression via cross-validation

Predictor	B	SE B	t-value	95% CI OR
ICU	1.342	0.124	10.781	3.83 (3,4.88)
Albumin_AVG	-0.473	0.106	-4.472	0.62 (0.51,0.77)
log_LDH	0.575	0.150	3.845	1.78 (1.33,2.38)
Potassium_AVG	0.482	0.105	4.578	1.62 (1.32,1.99)
Hypertension	0.510	0.100	5.104	1.66 (1.37,2.02)
Chloride_AVG	-0.053	0.009	-6.024	0.95 (0.93,0.97)
BloodPressureDiastolicNBR	-0.092	0.016	-5.845	0.91 (0.88,0.94)
Bicarbonate_AVG	-0.048	0.014	-3.385	0.95 (0.93,0.98)
TemperatureFahrenheitNBR	0.150	0.042	3.547	1.16 (1.07,1.26)
log_Glucose	0.420	0.137	3.069	1.52 (1.16,1.99)
Calcium_AVG	-0.345	0.097	-3.550	0.71 (0.59,0.86)
age	0.006	0.003	2.292	1.01 (1,1.01)
log_ANEU	1.080	0.280	3.853	2.94 (1.7,5.1)
log_WhiteBloodCellCNT	-0.837	0.260	-3.216	0.43 (0.26,0.72)
CRPT_status	-0.256	0.097	-2.654	0.77 (0.64,0.94)
RespirationRateNBR	0.023	0.009	2.679	1.02 (1.01,1.04)
sex	0.196	0.093	2.119	1.22 (1.01,1.46)
INR_status	0.216	0.112	1.936	1.24 (1,1.54)
BloodPressureSystolicNBR	-0.032	0.008	-3.877	0.97 (0.95,0.98)

The optimal probability cutoffs determined by Youden's Index for the logistic regression, neural network, and random forest were 0.283, 0.289, and 0.296 respectively. The random forest produced an average accuracy of 0.759, sensitivity of 0.653, specificity of 0.791, AUROC of 0.796, and AUPRC of 0.505 on the holdout test sets. The top 10 most important features included in the random forest model were lactate dehydrogenase, glucose, calcium, albumin, heart rate, age, chloride, potassium, bicarbonate, absolute neutrophil count and are shown in Figure A12.

The logistic regression produced an average accuracy of 0.766, sensitivity of 0.642, specificity of 0.804, AUROC of 0.775, and AUPRC of 0.501 on the holdout test sets. Predictors that were highly associated with increased risk of AKI were lactate dehydrogenase (OR=1.78; 95% CI 1.33-2.38), ICU on admission (OR=3.83; 95% CI 3-4.88), pre-existing hypertension (OR=1.66; 95% CI 1.37-2.02), increases in potassium (OR=1.62; 95% CI 1.32-1.99). The aggregated accuracy across the 5 imputed test sets for stepwise logistic regression via cross-validation with interactions was 0.793 and produced AUROC of 0.752.

The neural network produced an average accuracy of 0.772, sensitivity of 0.643, specificity of 0.811, AUROC of 0.787, and AUPRC of 0.505 on the holdout test sets. The ROC and Precision-Recall curves produced for each prediction model are shown below in Figures 6 and 7 respectively. Model calibration is shown in Figure 8 showing the relationship between the average predicted probabilities of observations in each bin and the true proportion of AKI in each bin. The Hosmer-Lemeshow Test statistics to assess goodness of fit for the respective models are shown in Table 4

Figure 6: ROC and for RF, NN, and LR

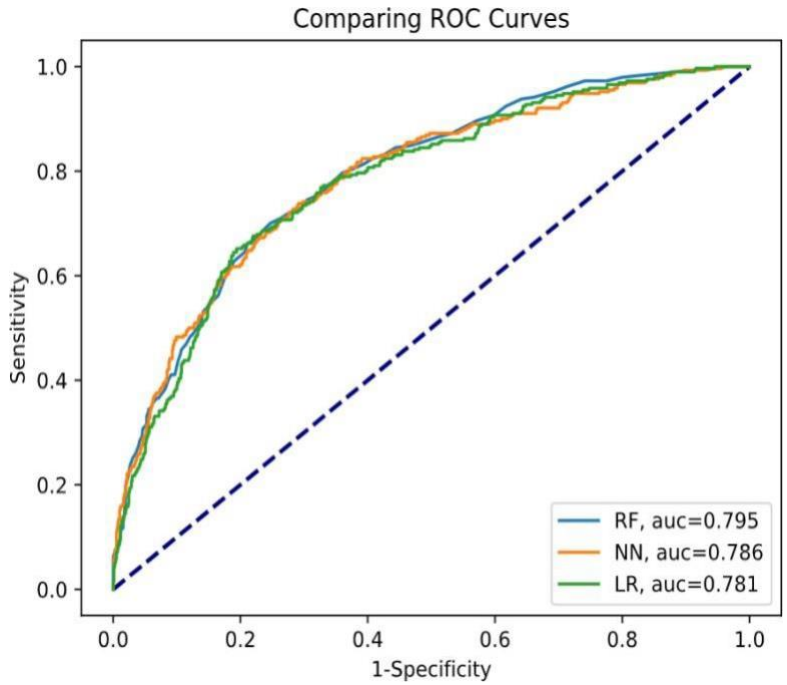


Figure 7. Precision Recall curves for RF, NN, and LR

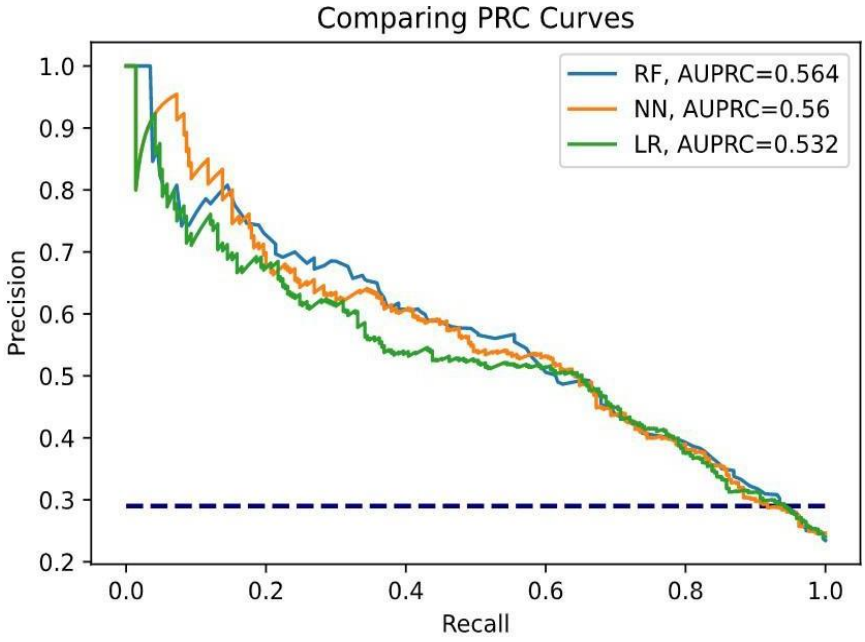


Figure 8. Model Calibration for RF, NN, and LR

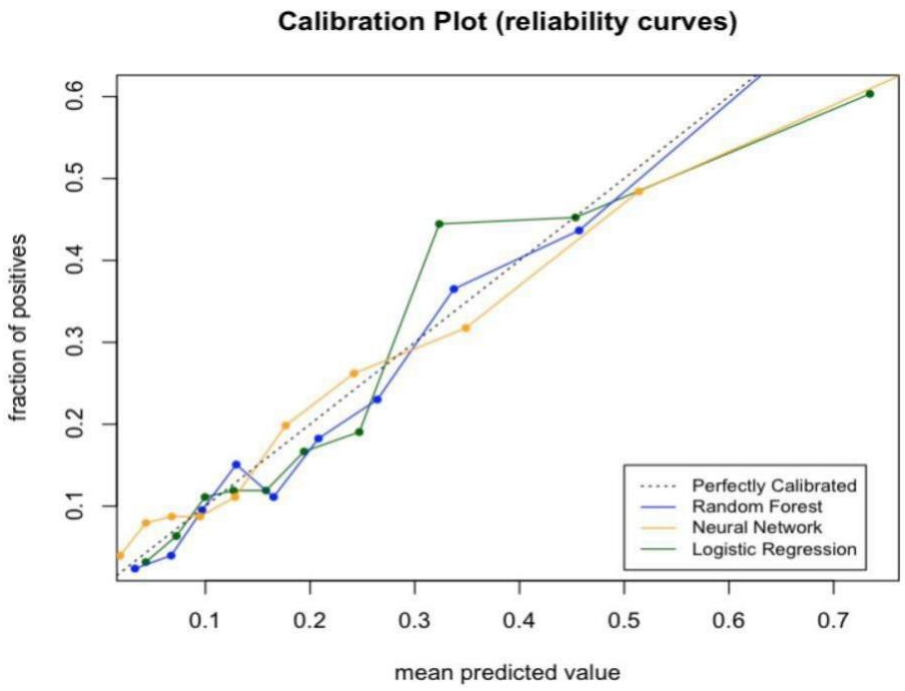


Table 4. Hosmer-Lemeshow Chi-Squared Test

Model	X-Squared	DF	P-value
Logistic Regression	23.07	8	0.0033
Random Forest	7.96	8	0.4377
Neural Network	25.17	8	0.0015

Effect plots for the predictors included in the main effects logistic regression model are shown in Figure A8**.** The effect plots show the relationship between predictor variables and the probability of AKI predicted from the logistic regression holding other variables constant.

Residual diagnostic plots for the main effects logistic regression model are shown in Figures A9, A10, and A11 respectively. Figure A9 shows the residuals plotted against predictors. The residuals are approximately randomly distributed suggesting the assumption of a linear relationship between predictors and AKI is appropriate. Figure A10 shows the binned residual plot for the logistic regression where data are grouped into bins based on

predicted probabilities and then plotted against the average residuals within each bin. The points lying within the 95% confidence bands suggests the model is well specified. Figure A11 shows Cook's Distance to identify influential points and potential outliers that may negatively affect the logistic regression model. The two labeled points (index=336 and index=2039) correspond to the observations furthest away from the average. There are few observations with large Cook's distance relative to other observations suggesting outliers are likely not influencing model results. Comparisons of predicted probabilities for 5 individuals are shown for the three models in Figure A13. There is little variation in predicted probabilities from the models for individuals found to be at low risk of AKI. For individuals identified to be at a greater risk of AKI there is much more variation between predicted probabilities from the three models.

4. Discussion

The previous sections describe the development of logistic regression, random forest, and neural network models to predict Acute Kidney Injury (AKI) as a binary outcome in hospitalized COVID-19 patients. Electronic Medical Records of 5,033 hospitalized COVID-19 patients at Brigham & Women's and Massachusetts General Hospital were investigated retrospectively to develop these models using patient's information available at or soon after admission including pre-existing medical conditions, vitals, lab results, and demographics.

The random forest produced the strongest results of three learning algorithms based on highest AUROC of 0.795 and highest sensitivity of 0.653 indicating it was the best of the three models at correctly identifying patients who experienced AKI. Additionally, the random forest was the only model that passed the Hosmer-Lemeshow Goodness of Fit test of calibration. The logistic regression (AUROC=0.781) and neural network

(AUROC=0.786) produced very comparable results to that of the random forest and yielded higher accuracy and specificity. However, both the neural network and logistic regression had poor model calibration based on the Hosmer-Lemeshow test statistics and were overpredicting for higher risk individuals. It is possible that potential nonlinearities are not appropriately being accounted for in the logistic regression and neural network causing poor calibration, where the random forest inherently takes nonlinearity into account. The sensitivity for all three models leave room for improvement especially for a clinical application where the true positive rate is prioritized. Decreasing the probability classification thresholds for the three models will improve model sensitivity at the expense of lower specificity, and will improve the respective model's ability to correctly identify patients that experience AKI.

Many features were found to have a statistically significant impact on the odds of AKI occurrence as reported from the stepwise logistic regression via cross-validation. The association of these predictors with AKI are consistent with findings from other studies investigating these relationships. For example, the increased odds of AKI based on ICU admission agrees with findings from study using COVID-19 data from MEDLINE, Embase, the Cochrane Library, and a registry of preprinted studies to investigate the prevalence of AKI in hospitalized COVID-19 patients, where 46% (95% CI, 35%-57%) of COVID-19 patients admitted to the ICU experienced AKI [28]. Additionally, a study at Mount Sinai found higher levels of potassium (OR 1.3; 95% CI 1.07-1.6) to be associated with greater risk of AKI [12]. Surprisingly, there were no significant interactions deemed from logistic regression via cross-validation. An interaction between hypertension and blood pressure may have been appropriate where the effect of blood pressure on AKI is influenced by the presence of hypertension. A possible explanation for this is that patients with pre-existing

hypertension may be on medications to regulate their blood pressure that are not being taken into account.

4.1 Limitations

The model results should be interpreted with the following limitations in mind. Missing data were a prominent issue where 3.2% of patients were missing serum creatinine lab values upon admission. Multiple imputation by chained equations (MICE) was used to impute missing values introducing additional variability into the model predictions. Rubin's rule was implemented to aggregate the model results across 5 imputed training and test sets. Using a greater number of imputed training and test sets from MICE could result in more precise estimates for the fitted models and results. Additionally, there was no limit on the time interval at which AKI occurred during a patient's hospital stay. Since the models only included information available upon hospital admission, it is likely that AKI occurrences that occurred later in a patient's hospital stay are not represented well. Making the predictions of AKI time sensitive may improve model performance. The neural network and random forest may not have optimal performance because only a subset of possible hyperparameters were tuned during the cross-validation process due to excessive computational time. The model findings are generalizable only to adults (>18) diagnosed with positive or presumptive COVID-19 that were hospitalized for at least 24 hours.

For practical purposes, the models developed require many input features to make predictions, and therefore may not be straightforward to implement in practice if not all data is available. Additionally, one of the strongest predictors of AKI is whether or not a patient was admitted to the ICU at or soon after admission. It is possible that patients will not be admitted to the ICU on admission due to availability of rooms rather than the severity of their illness.

4.2 Future Work

Further model validation has yet to be performed on an external dataset and needs to be investigated before implementing these prediction models in practice. Additionally, the use of a staging system ranging from stage 1 (least severe) to stage 3 (most severe) of AKI would be more informative when determining the appropriate triage of individuals. Implementing ordinal regression and multiclass prediction algorithms could be used to predict the stage of AKI and may be more appropriate for practical use.

Alternative regression approaches such as penalized lasso regression may lead to a simpler model with fewer predictors and should be investigated to see if more accurate results can be obtained. Additionally, more robust algorithms such as XGboost have been found to produce superior prediction results to logistic regression, random forest, and neural networks in many healthcare applications and could improve prediction accuracy.

5. Conclusion

This study found that logistic regression, random forest, and neural network machine learning algorithms can effectively be used to predict AKI occurrence in hospitalized COVID-19 patients using information available upon admission. Overall, the random forest classifier produced the strongest prediction and calibration results when evaluated on the holdout test sets with an aggregated AUROC of 0.795 and chi-squared goodness of fit statistic of 7.96 (p-value=0.4377).

Appendix

Table A1 : Summary Statistics of Predictors

Variable Name	Normal Reference Range	Mean	SD	Min	25%	50%	75%	Max	% Missing
Potassium	3.6-5.1 mmol/L	4.04	0.44	2.20	3.75	4.00	4.30	7.10	3.79
Calcium	8.9-10.3 mg/dL	8.80	0.55	6.60	8.47	8.80	9.13	12.72	3.76
Chloride	98-107 mmol/L	101.32	5.08	72.67	98.33	101.25	104.00	136.00	3.76
Bicarbonate	23-29 mmol/L	23.83	3.31	9.00	22.00	24.00	25.67	45.17	3.74
Albumin	3.5-5.2 g/dL	3.49	0.53	1.30	3.16	3.50	3.83	5.40	8.56
WBC	4-11 K/uL	7.79	4.10	0.15	5.12	6.86	9.35	48.11	2.01
Glucose	70-115 mg/dL	139.35	57.46	50.50	104.00	119.67	150.78	552.50	3.70
LDH	140-280 U/L	329.99	154.90	86.00	226.00	294.58	392.08	2031.50	24.74
ANEU	1.5-8 K/uL	5.77	3.69	0.00	3.37	4.82	7.04	35.66	5.84
Age	>=18 years	62.32	19.07	18.09	48.85	63.83	77.35	107.18	0.00
SBP	90-120 mmHg	124.72	20.59	0.00	111.00	123.00	137.00	214.00	0.10
DBP	60-80 mmHg	69.64	11.84	0.00	61.00	69.00	77.00	140.00	0.10
Body Temperature	97-99 Fahrenheit	98.03	0.99	89.60	97.50	98.00	98.50	106.90	0.16
Heart Rate	60-100 Beats Per Minute	79.07	19.95	0.00	69.00	79.00	90.00	166.00	0.10
BMI	18.5 to <25 kg/m^2	28.92	7.23	8.90	24.00	28.00	32.51	72.78	1.09
Respiration Rate	12-20 Breaths Per Minute	19.23	4.67	0.00	18.00	18.00	20.00	60.00	0.26
Sex	Male=1,Female=0	0.52							0
ICU Admission	Admitted=1, else 0	0.12							0
INR	Received Test=1, else 0	0.71							0
CRPT	Received Test=1, else 0	0.48							0
PT	Received Test=1, else 0	0.71							0
Hypertension	Yes=1, else 0	0.38							0
Liver Disease	Yes=1, else 0	0.01							0
Peripheral Vascular Disease	Yes=1, else 0	0.02							0

Figure A1. Distributions of Continuous Variables

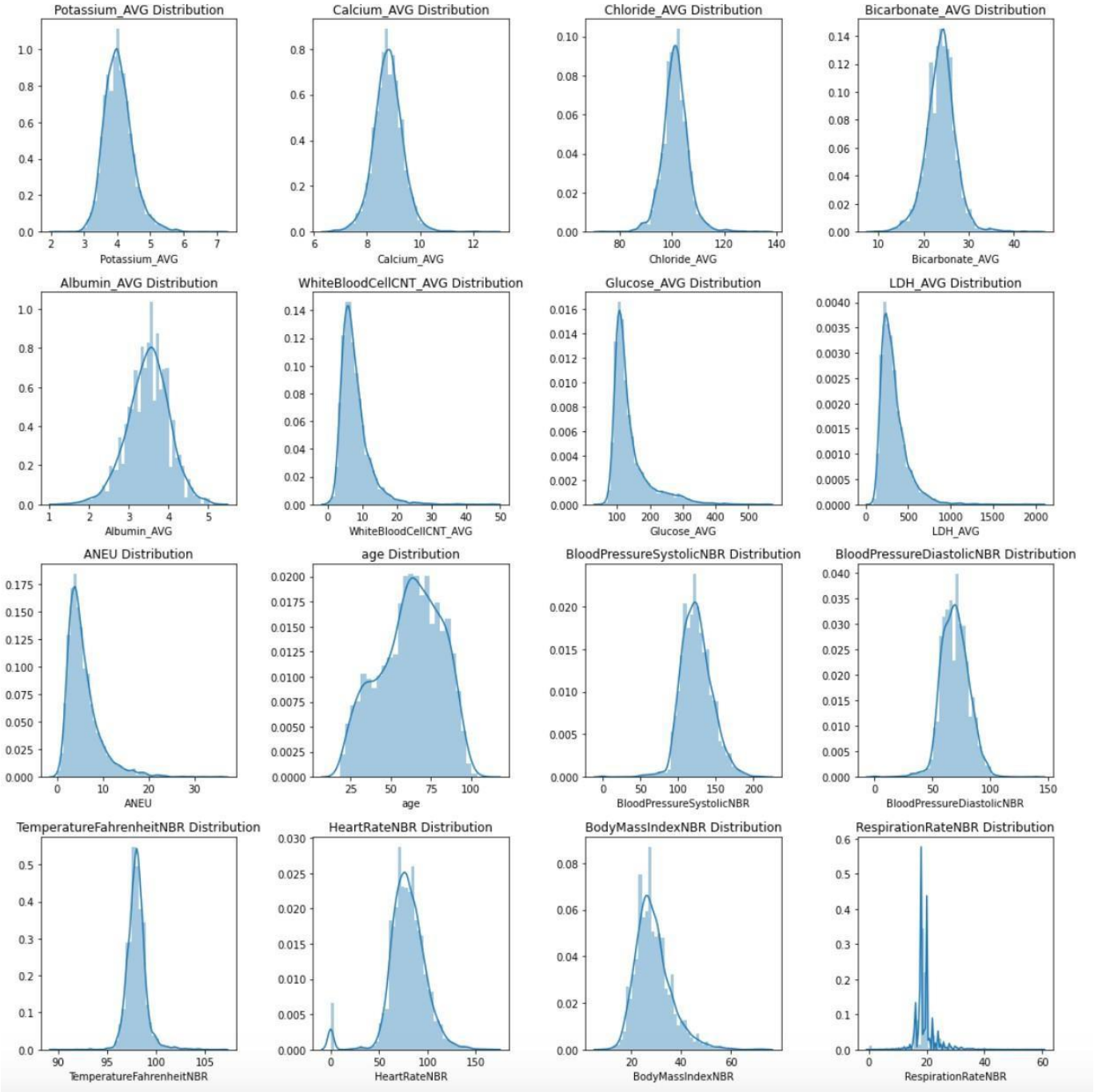


Figure A2. Distributions of Log Transformed Variables

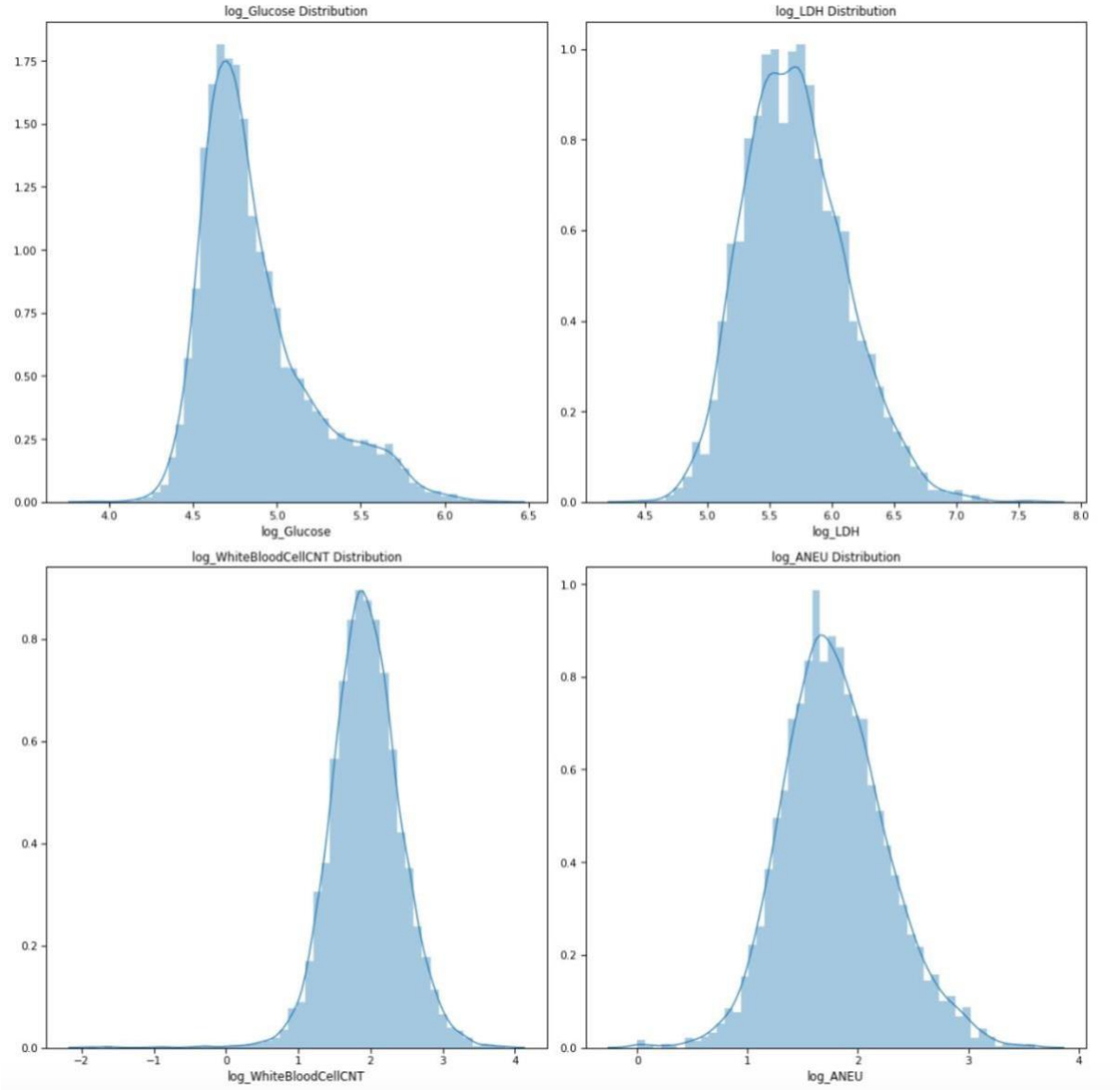


Figure A3. Proportions of missing data within each variable

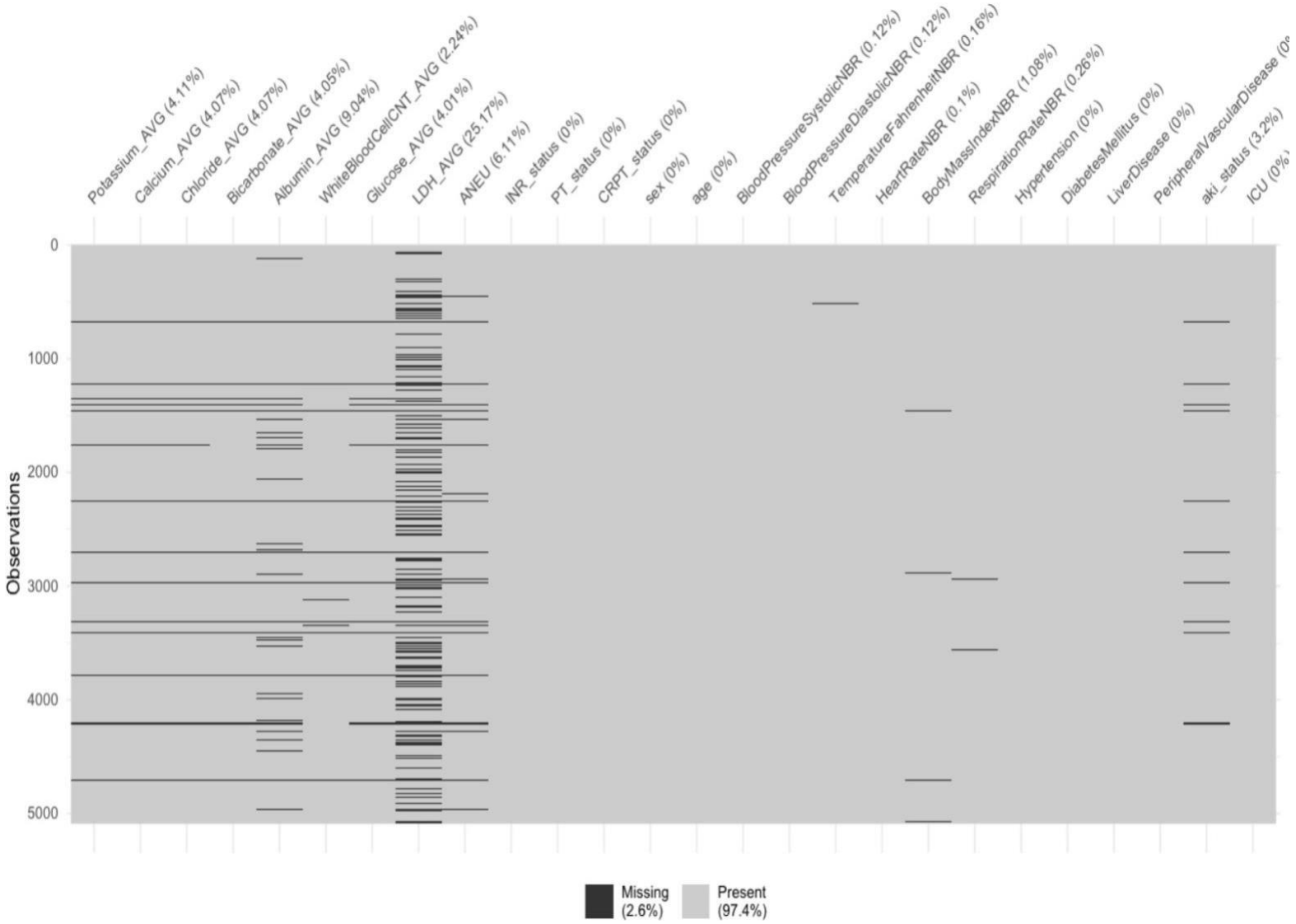


Figure A4. Pairwise comparison of variable distributions when missing and not missing

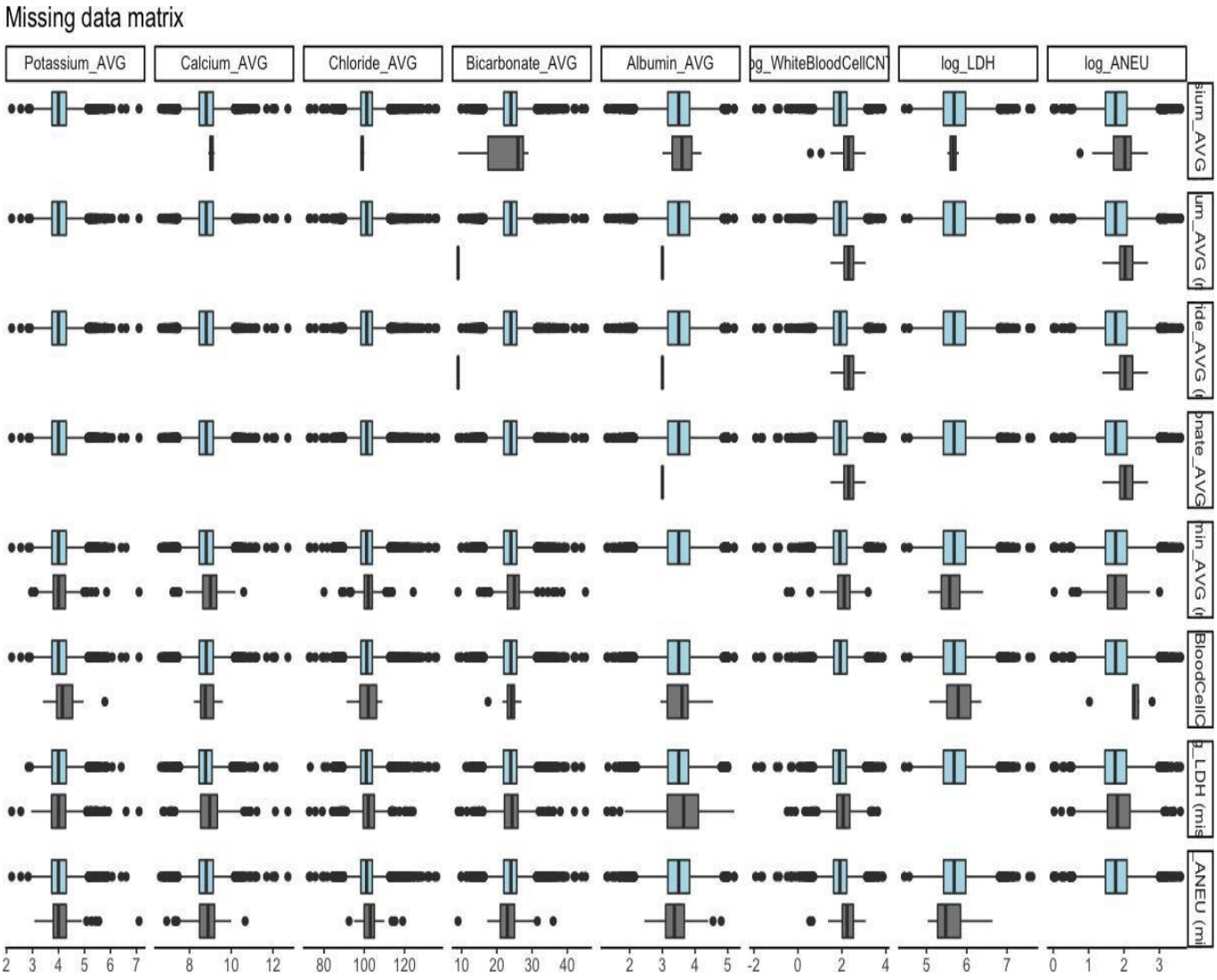


Figure A5. MICE Imputed Distributions vs Observed Value Distributions in Training Set

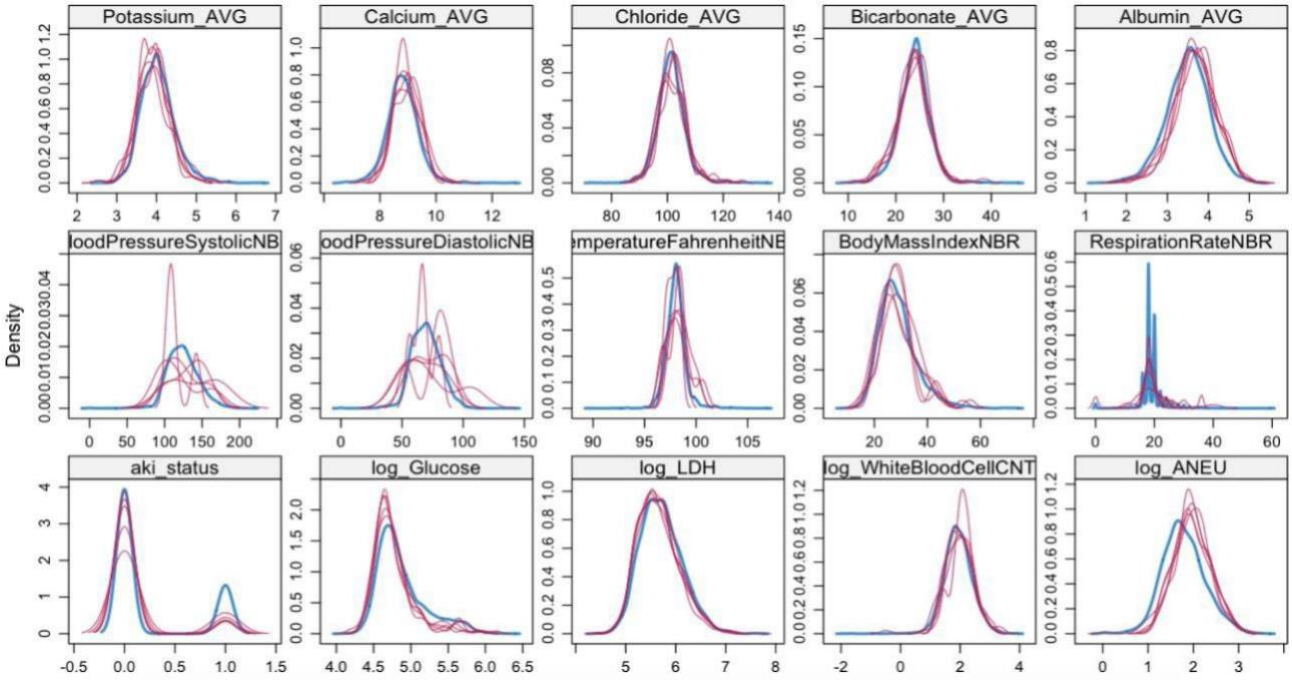


Figure A6: MICE Imputed Distributions vs Observed Value Distributions in Testing Set

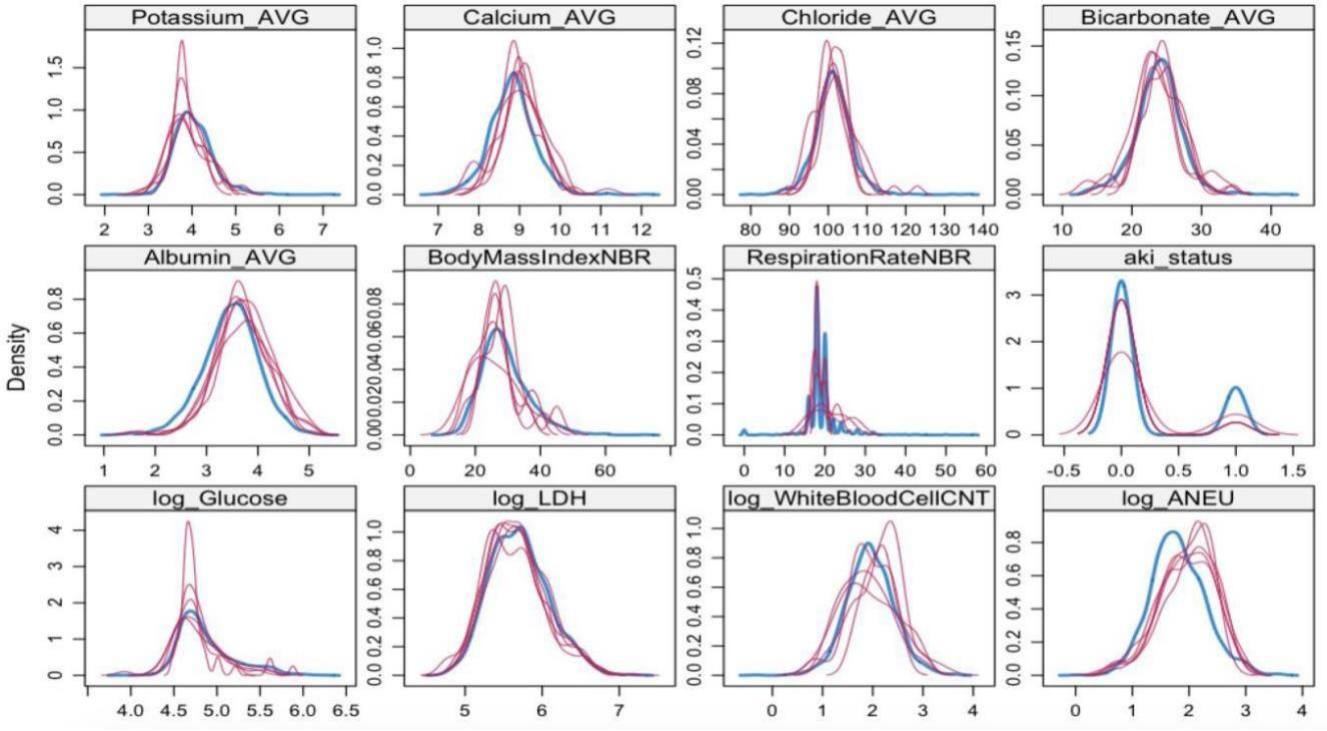


Table A2. Incidence of AKI within subgroups

Subgroup	AKI	No AKI
Sex		
Male	27.28%	72.72%
Female	20.46%	79.54%
ICU		
Admitted to ICU (within first 48 hours)	59.64%	40.36%
Not Admitted to ICU (within 48 hours)	19.11%	80.89%
INR_status		
Recorded INR test	28.03%	71.97%
no recorded INR Test	14.13%	85.87%
PT_status		
Recorded PT	28.00%	72.00%
no recorded PT	14.16%	85.84%
CRPT_status		
Recorded CRPT	25.00%	75.00%
no recorded CRPT Test	23.09%	76.91%
Liver Disease		
pre-existing liver disease	0.46%	99.54%
no pre-existing liver disease	45.46%	54.54%
Hypertension		
pre-existing hypertension	30.60%	69.40%
no pre-existing hypertension	20.01%	79.99%
Peripheral Vascular Disease (PVD)		
pre-existing PVD	35.00%	65.00%
no pre-existing PVD	23.73%	76.27%

Figure A7. Confusion Matrices for LR, RF, and NN

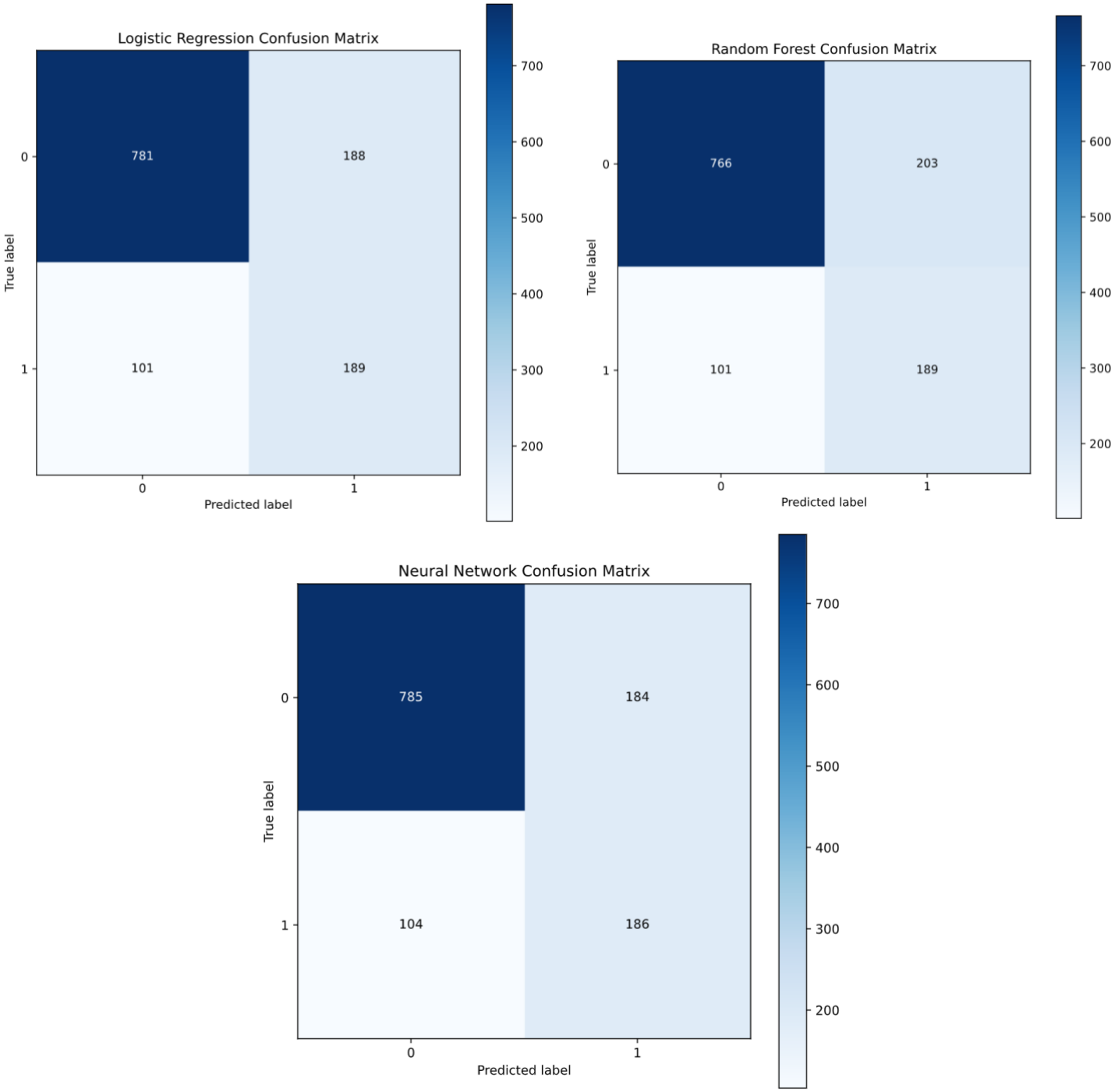


Figure A8. Effects Plots for LR

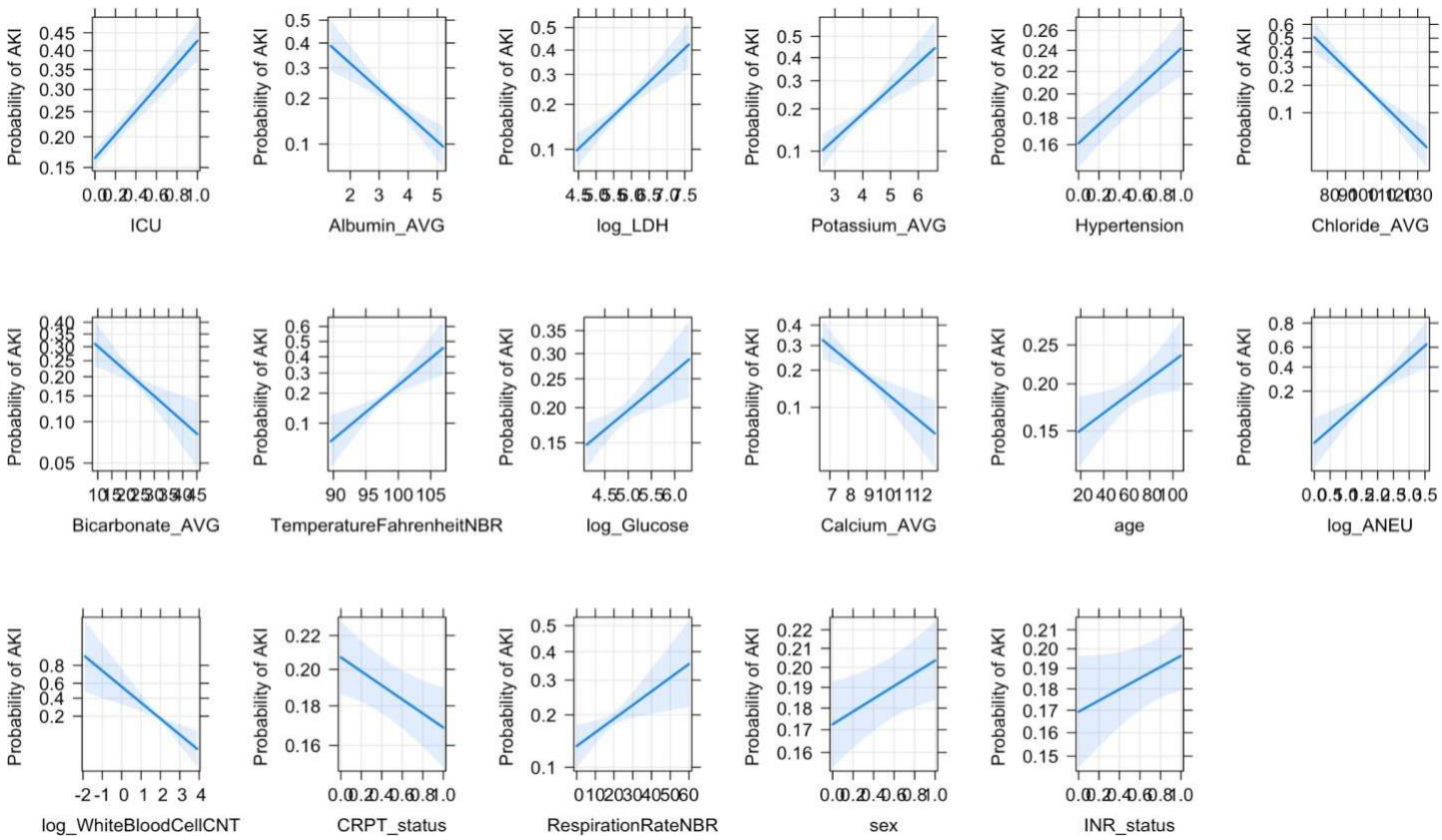


Figure A9. Residuals vs Predictors from LR

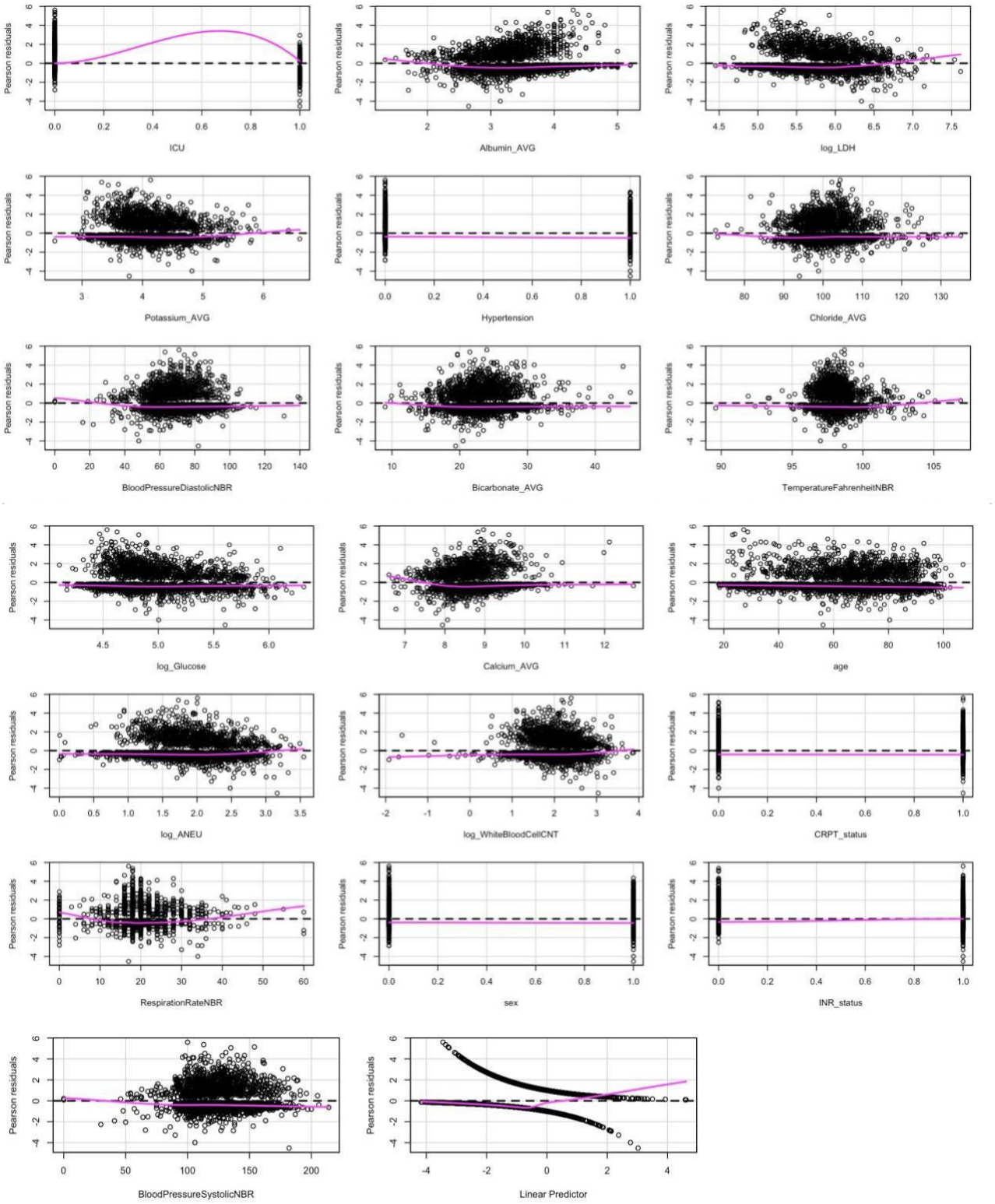


Figure A10. Logistic Regression Binned Residual Plot

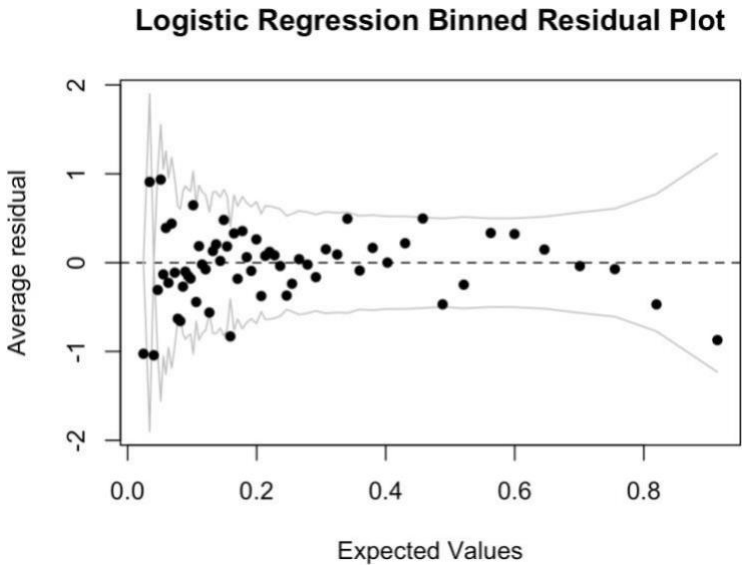


Figure A11 Cook's Distance for Logistic Regression

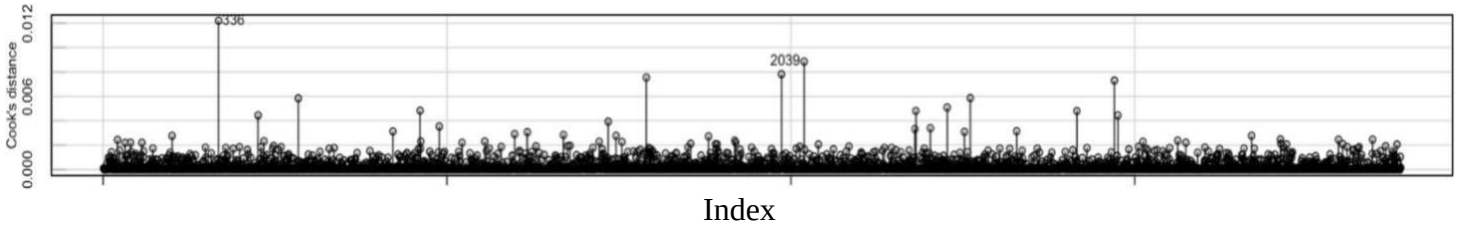


Figure A12. Random Forest Feature Importance

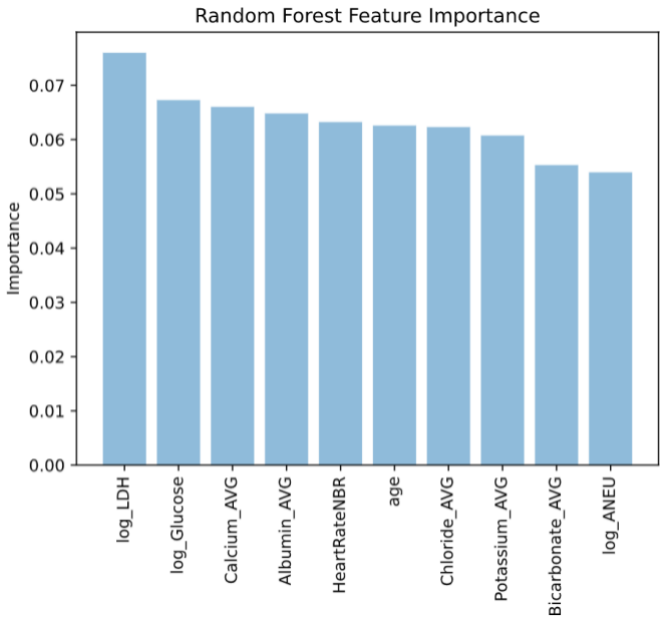
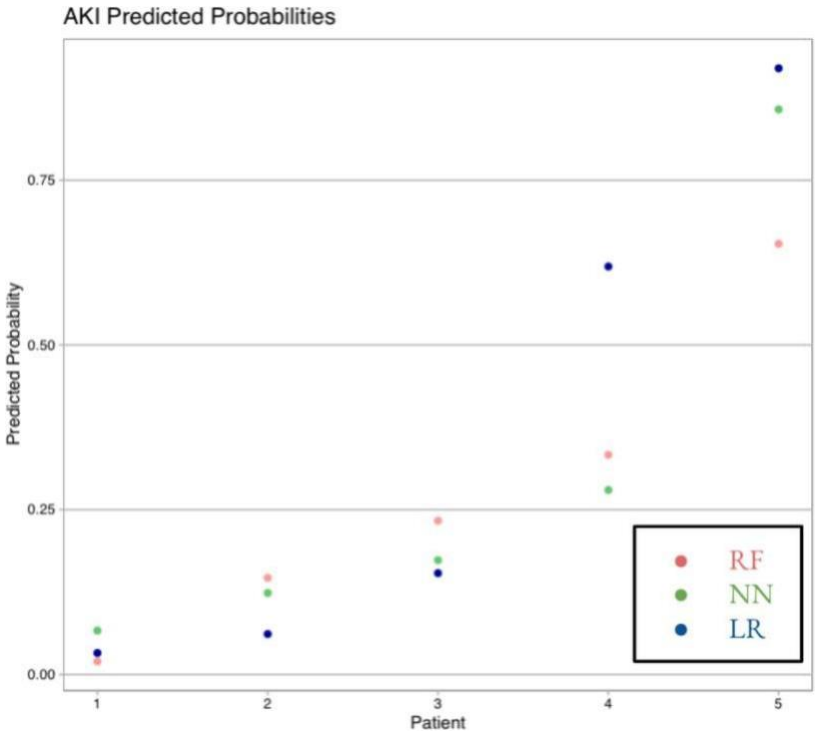


Figure A13. Comparison of Model Predicted Probabilities for Logistic Regression, Random Forest, and Neural Network



References

1.) Coronavirus Cases: *Worldmeter*, www.worldometers.info/coronavirus/.

2.) Makris, Konstantinos, and Loukia Spanou. "Acute Kidney Injury: Definition, Pathophysiology and Clinical Phenotypes." *The Clinical Biochemist. Reviews*, The Australian Association of Clinical Biochemists, May 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC5198510/.

3.) Fornell, Dave. "The Cardiovascular Impact of COVID-19." *DAIC*, 28 July 2020,www.dicardiology.com/article/cardiovascular-impact-covid-19.

4.) "Acute Kidney Injury (AKI)." *National Kidney Foundation*, 30 Oct. 2020, www.kidney.org/atoz/content/AcuteKidneyInjury.

5.) "LitCovid - NCBI - NLM - NIH." *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/research/coronavirus/publication/33319190.

6.) Piper Julie Hughes, MD. *Classification Systems for Acute Kidney Injury: Background, RIFLE Classification, Acute Kidney Injury Network*, Medscape, 5 Jan. 2021, emedicine.medscape.com/article/1925597-overview.

7.) "Acute Kidney Injury and Chronic Kidney Disease - What's the Difference?" *NursingCenter*, www.nursingcenter.com/ncblog/january-2020/acute-kidney-injury-and-chronic-kidney-disease.

8.) "Acute Kidney Failure." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 23 July 2020, www.mayoclinic.org/diseases-conditions/kidney-failure/diagnosis-treatment/drc-20369053.

9.) "What Is Dialysis?" *National Kidney Foundation*, 30 Oct. 2020, www.kidney.org/atoz/content/dialysisinfo.

11.) Chan, Lili, et al. "Acute Kidney Injury in Hospitalized Patients with COVID-19." *MedRxiv*, Cold Spring Harbor Laboratory Press, 1 Jan. 2020, www.medrxiv.org/content/10.1101/2020.05.04.20090944v1.

12.) Vaid1, A., Somani1*, S., Russak1, A., Freitas1, J., Chaudhry1, F., Paranjpe1, I., . . . Authors..., L. (n.d.). Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients With COVID-19 in New York City: Model Development and Validation. Retrieved December 11, 2020, from https://www.jmir.org/2020/11/e24018.

13.) "Youden Index." *Youden Index - an Overview | ScienceDirect Topics*, www.sciencedirect.com/topics/medicine-and-dentistry/youden-index.

14.) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R.* New York: Springer.

15.) Alex. (2020, May 12). Feedforward Neural Networks and Multilayer Perceptrons. Retrieved December 11, 2020, from https://boostedml.com/2020/04/feedforward-neural-networksand-multilayer-perceptrons.html.

16.) Bagheri, R. (2020, August 28). An Introduction to Deep Feedforward Neural Networks. Retrieved December 11, 2020, from https://towardsdatascience.com/an-introduction-to deep-feedforward-neural-networks-1af281e306cd

17.) Kingma, Diederik, and Jimmy Ba. "Adam: A Method for Stochastic Optimization." *ArXiv.org*, 22 Dec. 2014, arxiv.org/abs/1412.6980v1.

18.) Cheng Y, Luo R, Wang K, et al. Kidney disease is associated with in-hospital death of patients with COVID-19. 2020 March 5

19.) Batlle, Daniel, et al. "Acute Kidney Injury in COVID-19: Emerging Evidence of a Distinct Pathophysiology." *American Society of Nephrology*, American Society of Nephrology, 1 July 2020, jasn.asnjournals.org/content/31/7/1380.

20.) Saito, Takaya, and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLOS ONE*, Public Library of Science, journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0118432.

21.) Azur, Melissa J., et al. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" *International Journal of Methods in Psychiatric Research*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/.

22.) Kirasich, Kaitlin; Smith, Trace; and Sadler, Bivin (2018) "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," *SMU Data Science Review*: Vol. 1 : No. 3 , Article 9.

23.) Dreiseitl, Stephan, and Melanie Osl. "Testing the Calibration of Classification Models from First Principles." *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, American Medical Informatics Association, 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3540450/.

24.) Brownlee, J. (2019, October 25). Difference between a batch and an epoch in a neural network. Retrieved March 30, 2021