6-2021

Developing Prediction Models for Kidney Stone Disease

Joseph Palko

Running Title: Developing Prediction Models for Kidney Stone Disease

Developing Prediction Models for Kidney Stone Disease

By

Joseph Palko

****

Submitted in partial fulfillment
of the requirements for
Honors in the Department of Mathematics

UNION COLLEGE

March, 2021

ABSTRACT

PALKO, JOSEPH   Developing Prediction Models for Kidney Stone Disease
Department of Mathematics, March 2021.

ADVISOR: [Professor Jue Wang]

Kidney stone disease has become more prevalent through the years, leading to high treatment cost and associated health risks. In this study, we explore a large medical database and machine learning methods to extract features and construct models for diagnosing kidney stone disease.

Data of 46,250 patients and 58,976 hospital admissions were extracted and analyzed, including patients' demographic information, diagnoses, vital signs, and laboratory measurements of the blood and urine. We compared the kidney stone (KDS) patients to patients with abdominal and back pain (ABP), patients diagnosed with nephritis, nephrosis, renal sclerosis, chronic kidney disease, or acute and unspecified renal failure (NCA), patients diagnosed with urinary tract infections and other diseases of the kidneys and the uterus (OKU), and patients with other conditions (OTH). We built logistic regression models and random forest models to determine the best prediction outcome.

For the KDS vs. ABP group, a logistic regression model using the five variables including age, mean respiratory rate, blood chloride, blood creatinine, and blood CO2 levels from the patients' first lab results gave the best prediction accuracy of 0.699. This model maximized sensitivity with a value of 0.726. For KDS vs. NCA we found that a logistic regression using the Elixhauser score and blood urea nitrogen (BUN) values from the first lab results for patients with first admittance produced the best outcome, with an accuracy of 0.883 and maximized specificity of 0.898. For KDS vs. OKU a logistic regression using the estimated glomerular filtration rate (EGFR) calculated from the average lab values gave the best outcome, with an accuracy of 0.852 and maximized specificity of 0.922. Finally, a logistic regression using age, EGFR, BUN, blood creatinine, and blood CO2 gave the best outcome for KDS vs. OTH, with an accuracy of 0.894 and maximized specificity of 0.903. This research gives the medical field models to potentially use on kidney stone patients. It also provides a steppingstone for researchers to build off if they want to build kidney stone models for a different population of patients.

Table of Contents

Appendix

1. Background

1.1 Kidney Stone Information

Kidney stones are hard deposits of salts and minerals that form in the kidney [Kidney Stones, 2020]. These stones vary greatly in size ranging usually from about 1mm to 10mm [IQWiG]. Most stones that are less than 5 millimeters in diameter travel through the bladder without any problems. Only about 50% of the stones that are between 5mm and 10mm leave the body without assistance. Stones that have a diameter greater than 10 millimeters usually need to be treated [IQWiG]. Kidney stones have a lot of different symptoms with intense pain being the most common. Other common symptoms include nausea and vomiting [Chen, et.al., 2018]. Kidney stone presence has increased slightly from a 5.2% prevalence to an 8.8% prevalence over an 18 year period from 1994 to 2012 [Chen, et.al., 2018]. The cost of treating kidney stones is extremely high, with it costing an estimated total of 4 billion dollars in 2012 [Chen, et.al., 2018]. People who have experienced a kidney stone are at greater risk of developing chronic kidney disease [Chen, et.al., 2018].

There are many different things that can increase a person's chances of developing kidney stones. These things include diet, liquid intake, and even the medicines that someone is taking. People who are obese, have diabetes, have higher BMI, have cholelithiasis, or have gout, are at a greater risk of developing kidney stones [Chen, et.al., 2018]. Other health factors that increase one's risk include hypercalciuria, high blood pressure, osteoporosis, kidney cysts, cystic fibrosis, parathyroid disease, inflammatory bowel disease, chronic diarrhea, and surgeries that affect the stomach or the intestine [Kidney Stones, (c)]. Males and whites are also more likely to develop kidney stones [Kidney Stones, (c)]. People who have a family history of kidney stones are at a greater risk of developing kidney stones [Kidney Stones, (b)]. Older patients are also more at risk of developing kidney stones, as kidney stones are more likely to develop during the ages of 40-60 [Kidney Stones, (b)]. People who have diets that consist of high protein and low carbs or diets with high sodium, have increased chances of developing a kidney stone [Kidney Stones, (c)]. People who do not drink very much liquid are more prone to kidney stones [Kidney Stones, (c)]. Some medications that make people more prone to getting kidney stones include diuretics, calcium based antacids, crixivan, topamax, dilantin, cipro, and ceftriaxonie [Kidney Stones, (c)].

1.2 Types and Formation of Kidney Stones

Kidney stone formation involves crystal nucleation, specifically heterogeneous nucleation. Nucleation is when ions, which are usually dissolved in a solvent, come together and form clusters [Khan, 2016]. Having low amounts of fluid intake decreases the necessary solvent, making it easier for ions to combine. Heterogeneous nucleation is the type of nucleation that requires a lower supersaturation [Khan, 2016]. In other words, a lower level of solute needs to be present for ions to start combining. There are four different types of kidney stones that can form. These types are calcium stones, struvite stones, uric acid stones, and cystine stones. Calcium stones are formed from calcium oxalate and calcium phosphate crystals [Kidney Stones, (a)]. These stones are also the most common type of kidney stone. Struvite stones form from increased production of ammonia and are called infectious stones [Khan, 2016]. Struvite stones usually grow very quickly and can get very large [Kidney Stones, (a)]. They are only seen in patients who frequently have urinary tract infections [Khan, 2016]. Uric acid stones consist of only about 10% of all kidney stone diseases and are found most in obese or insulin resistant people [Khan, 2016]. These stones are formed due to overly acidic urine and are also more common in people who have high protein diets [Kidney Stones, (a)]. Cystine stones are usually only found in people who have a family history of the disease called cystinuria [Kidney Stones, (a)]. This disease causes the kidneys to excrete too much of an amino acid, which then allows the amino acid to come together and form a kidney stone [Kidney Stones, (a)].

2. Kidney Stone Metrics

2.1 Estimated Glomerular Filtration Rate

The Estimated Glomerular Filtration Rate or EGFR refers to the filtration rate of the kidney and it gives an idea of how well one's kidney is functioning [Gilbert & Weiner, 2014]. A higher GFR value means a well-functioning kidney. Normal average GFR values are approximately 130 and 120 mL/min/1.73m^2 for young men and women, respectively [Gilbert & Weiner, 2014]. There are many EGFR formulas, but all of them use the following variables: a patient's serum creatinine level, age, sex, and ethnicity [eGFR, 2020]. One complication with these formulas is that they were created based on data from people who were white or black [eGFR, 2020]. This means that people of different ethnicities may not get good results from any of the formulas. Another issue with having so many different formulas is that two hospitals could get different GFR values [eGFR, 2020]. This could cause confusion for the patients. Table 1 shows some EGFR formulas and how they are calculated.

Table 1 Various formulas for calculating glomerular filtration rate (GFR)

Cockcroft-Gault Formula	Ccr (mL/min) = (140-age) × weight/72 × Scr × 0.85 [if female]
MDRD Study Equation for Use with Standardized Serum Creatinine (Four-Variable Equation)	GFR (mL/min/1.73m^2) = 175 × SCr^(−1.154) × age^(−0.203) × 0.742 [if female] × 1.210 [if black]
CKD-EPI Equation for Use with Standardized Serum Creatinine	GFR (mL/min/1.73m^2) = 141 × min(Scr/κ, 1)^α × max(Scr/κ, 1)^(1.209) × 0.993Age × 1.018 [if female] × 1.157 [if black]
Schwartz Formula (Younger than 18 Years of Age):	eGFR = 40.7 × [HT/Scr]^(0.640) × [30/BUN]^(0.202)

2.2 Elixhauser Comorbidity Index (ECI)

The Elixhauser Comorbidity Index was first created in 1998 and its goal is to summarize the overall disease burden that a patient has [ECI, 2021]. This index looks at 30 individual diseases and gives a point value based on a yes or no answer to the question of whether they have the disease. The scores for each individual question range from -7 to 12 points [ECI, 2021]. Some of the diseases that this index looks at are congestive heart failure, hypertension, paralysis, chronic pulmonary disease, renal failure, and liver disease [ECI, 2021]. The total scores for this index range from -19 to 89. The higher the score the greater the disease burden is for that individual [ECI, 2021]. In addition, a person who has a greater disease burden is also more likely to die in the hospital than a person with a low disease burden [ECI, 2021]. Other important indexes like the Charlson Comorbidity Index, Stone Score, and the research tool JESS, which stands for Joint Expert Speciation System are great tools when looking into kidney stones. We do not delve much into these topics, but additional information on these tools can be found in appendix 1.

3. Data Description

We want to investigate what factors seem to have the greatest influence in categorizing patients with kidney stones when compared to patients with kidney related diseases, urinary diseases, or other diseases in general. The data that we will be using in this project comes from the MIMIC-III Critical Care Database [MIMIC-III], free for approved researchers to use. This database contains 58,976 hospital admissions for the 46,520 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [MIMIC-III]. Medical information includes patient demographics, diagnoses, vital signs, laboratory measurements, admittance time, procedures done on the patient, medications that the patient is on, notes for patients, and so on.

We extracted adult (18 years and older) patients' information and divided them into five groups based on the ICD-9 diagnoses during hospitalization. The study groups are patients diagnosed with kidney stone (KDS), patients with abdominal and back pain (ABP), patients diagnosed with nephritis, nephrosis, renal sclerosis, chronic kidney disease, or acute and unspecified renal failure (NCA), patients diagnosed with urinary tract infections and other diseases of the kidneys and the uterus (OKU), and patients with other conditions (OTH). The extracted features are listed in table 2 below. In addition, we extracted all diagnoses for the patient and computed the EGFR and ECI for each patient.

The extracted data contained 534 KDS, 451 ABP, 16,701 NCA, 4,620 OKU, and 24,295 OTH patients. A subset of this larger data set was used to learn R programming as well as to get an understanding of some potential risk factors for kidney stones. This subset contained 534 KS patients and 445 ABP patients with only demographic information.

Table 2 Extracted features and ICD-9 codes

Demographics	Gender, ethnicity, age, height, weight, BMI
8 vital signs	Heart rate, systolic blood pressure, diastolic blood pressure, mean arterial blood pressure, respiratory rate, temperature Celsius, peripheral SpO2, glucose
52 lab blood measurements	Anion gap, albumin, bands, base excess, carbonate, total bilirubin, total calcium, chloride, calculated total co2, creatinine, creatine kinase, glucose, hematocrit, hemoglobin, lactate, lactate dehydrogenase (ld), lipase, magnesium, oxygen saturation, ph, phosphate, platelet count, po2, potassium, ptt, inr(pt), pt, red blood cells, red cell distribution width (rdw), sodium,urea nitrogen (BUN), uric acid, white blood cells, alkaline phosphatase, alanine aminotransferase (alt), asparate aminotransferase (ast), ammonia, anti-nuclear antibody, cholesterol ratio (total/hdl), hdl cholesterol, ldl cholesterol, total cholesterol, cortisol, d-dimer, folate, globulin, mch, mchc, mcv, thyroid stimulating hormone, triglycerides, vitamin B12
21 lab urine measurements	Ammonium, amylase, bilirubin, calcium oxalate crystals, creatinine, urine cysteine crystals, epithelial cells, glucose, ketone, leukocytes, magnesium, nitrite, ph, phosphate, protein, RBC, sodium, uric acid, urine volume, wbc, yeast
KDS ICD-9 codes	5920, 5921, 5929, 5940, 5941, 5942, 5948, 5949, 7880, V1301
APB ICD-9 codes	7890, 78900, 78901, 78902, 78903, 78904, 78905, 78906, 78907, 78909, 78960, 78961, 78962, 78963, 78964, 78965, 78966, 78967, 78969, 7242
NCA ICD-9 codes	5800, 5804, 58081, 58089, 5809, 5810, 5811, 5812, 5813, 58181, 8189, 5819, 5820, 5821, 5822, 5824, 58281, 58289, 5829, 5830, 5831, 5832, 5834, 5836, 5837, 58381, 58389, 5839, 587, 585, 5851, 5852, 5853, 5854, 5855, 5856, 5859, 7925, V420, V451, V4511, V4512, V560, V561, V562, V5631, V5632, V568, 5845, 5846, 5847, 5848, 5849, 586
OKU ICD-9 codes	5880, 5881, 5888, 58881, 58889, 5889, 5890, 5891, 5899, 591, 5930, 5931, 5932, 5933, 5934, 5935, 5936, 5937, 59370, 59371, 59372, 59373, 59381, 59382, 59389, 5939, 03284, 59000, 59001, 59010, 59011, 5902, 5903, 59080, 59081, 5909, 5950, 5951, 5952, 5953, 5954, 59581, 59582, 59589, 5959, 5970, 59780, 59781, 59789, 59800, 59801, 5990
OTH ICD-9 codes	All other ICD-9 codes that are not listed above

4. Methods

We first introduce some important models and quantifications used in data analysis and machine learning, including correlation, decision tree, random forest, logistic regression, odds ratio, confidence interval, and confusion matrix.

4.1 Correlations

Correlations are very important when trying to determine if two variables are dependent or independent. There are three types of tests that will tell us how much two variables are related to each other. These three tests are the Pearson correlation test, the chi square test, and Matthew's correlation coefficient (MCC). The Pearson correlation is used to measure the strength and direction of the relationship between two variables [Bradburn, 2020]. A key distinction for this test is that it can only be used with numerical variables. The values range from -1 to 1, with -1 meaning a strong negative correlation and 1 representing a strong positive correlation [Bradburn, 2020]. A value of 0 means that there is no relationship between the two variables. A strong negative correlation means as one variable increases in value the other variable will decrease in value. A strong positive correlation means as one variable increases in value, the other variable will also increase in value. Along with a correlation coefficient, this test gives a p-value. This determines the significance of the result. If the p-value is less than 0.05 then the correlation between the two variables is most likely not zero. If the p-value is greater than 0.05 then there is a possibility that the correlation could be zero. The chi square test sees how well the observed values for a given distribution fits with the distribution when the variables are independent [Agrawal, 2020]. This test can only be used with variables that are categorical [Agrawal, 2020]. This test will also give you a p-value. With a p-value of less than 0.05 we would conclude that the two variables are dependent. Otherwise, we would say they are independent. The MCC is a correlation coefficient that represents the correlation between the observed and predicted binary classifications [Fortney, 2018]. This test gives values that range from -1 to 1, like the Pearson correlation. Ideally, we would like to see a value of 1, as this means the model predicted perfectly. A value of -1 means the model did not predict correctly and predicts the opposite of what it should be. A value of 0 essentially means that the model is just as good as random assignment [Fortney, 2018].

4.2 Decision Trees and Random Forests

A decision tree is essentially a flow chart in the shape of a tree, where every node or "branch" of the tree answers a question that allows you to get closer and closer to a prediction [Decision Tree]. These decision trees aid in categorizing things like if someone is sick or not sick. The random combination of a bunch of decision trees is called a random forest. Random forests add more criteria and questions to the model, and it makes the predicted outcome more accurate than an individual decision tree [Donges, 2019]. An additional benefit of using random forests is that they usually don't over fit the data, unlike decision trees, which are more likely to over fit the data [Donges, 2019]. However, if a random forest gets too big and has too many branches, the time to conclusion could be very long, and thus not useful for real time decisions [Donges, 2019]. A random forest calculation is done using a term called bagging. This is basically taking a random sample of individual decision trees, to form a random forest, and training each of them with a data set [Krishni, 2019]. This process is repeated as many times as one requests, and a summarizing confusion matrix is presented. A data point is categorized by the amount of times it was predicted in a certain category. If it was predicted in the positive category more times than the negative category, then it will be assigned to the positive category in the summary matrix [Starmer]. We talk more about decision matrices and related terms in section 4.5. In addition, an OOB or out of the bag error rate is calculated based on these random forests. This error rate tells us which out of bag samples were incorrectly identified by the random forests. So, $(1 - OOB) * 100$ tells us the amount of out of bag samples that were correctly identified by the random forests [Starmer]. All the random forests in this paper use the Breiman’s Random Forest Machine Learning Algorithm. This algorithm uses in-bag and out-of-bag samples. These samples are created by the bootstrapping technique [Livingston, 2005]. Each time this is done ⅓ of the data is part of the out-of-bag sample. The other 2/3s of the data is used as the in-bag sample. This is repeated many times and the overall results are averaged [Livingston, 2005]. Using this method nullifies the need for cross validation.

4.3 Logistic Regression

An additional way to categorize things is to use logistic regression. This is different from linear regression as it predicts whether something is true or false instead of predicting a value [Starmer]. This type of regression fits an s-shaped curve that ranges from a value of 0 to a value of 1 [Starmer]. Linear regression uses numerical variables, but logistic regression uses continuous or categorical variables. One key aspect of logistic regression is that it assumes all variables are independent of each other, even though it may not be true [Logistic Regression]. Some outputs that come from the logistic regression include a coefficient for the variable, a p-value for this coefficient, and an intercept value. These values can then be used to make an equation that can be used to determine the log odds, which we discuss further in section 4.4. The intercept value indicates the initial log odds value, without other variables. The coefficient refers to the weight or the impact of a variable on the log odds. A logistic regression has the standard equation:

$$log(odds) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_k x_k \tag{1}$$

The p-value that corresponds to this coefficient tells whether the variable is significant to the model. If the p-value is less than 0.05 the variable is significant to the calculation of the odds ratio. If the p-value is greater than 0.05 the variable is not significant, which basically means that there is a possibility that the value of the coefficient could be zero. If this is the case the variable has no effect on the log odds and thus is deemed not significant.

Having these different types of models is good, but we need a way to test and see how good they are. A way to see how accurate a logistic regression model is, is to use AUROC, or area under the receiver operator characteristic (ROC). A ROC graph plots the sensitivity on the y-axis and 1-specificity on the x-axis [Grace-Martin, 2018]. Sensitivity tells us what percentage of patients who have a disease were correctly identified [Starmer]. Specificity tells us what percent of patients who don’t have a disease were correctly identified [Starmer]. We talk more in depth about sensitivity and specificity in section 4.5. It is up to the scientist which one they think is more important to maximize for their data. In our case, since we are dealing with data from the medical field, we feel it is more important to maximize our sensitivity. The AUC (area under the curve) ranges from a value of 0 to a value of 1. Ideally, we would want an AUC with a value of 1 as this tells us the model is perfect in categorizing patients. An AUC value of 0 is bad because it means the model is flipping the categorization [Narkhede, 2019]. That is, things that should be zeros are classified as 1’s by the model and vice versa. A benefit of using AUC is that it can be used to compare multiple ROC graphs [Starmer].

4.4 Odds Ratio and Confidence Intervals

The odds ratio and the log of odds ratio illustrate a relationship between two variables [Starmer]. To understand the odds ratio, we need to understand what is needed to calculate it. The odds ratio (or) is given by or=(a*d)/(b*c). Assuming A represents one condition and B represents another condition, a represents the number of times both A and B are present. b represents the number of times A is present, but B is absent. c represents when A is absent, but B is present. Finally, d represents when both A and B are absent [Odds Ratio]. The log odds ratio is given by

ln(or) = ln(a*d/b*c) [Odds Ratio].

To find the confidence interval for the odds ratio, the standard error of the log of odds ratio is used. The standard error of the log of odds ratio is given by

$SE(ln(or)) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$ [Comparing Frequencies] **(2)**

Then to finally calculate the confidence interval for the odds ratio we use the equation

$e^{(ln(or)\pm[1.96*SE(ln(or))])}$[Comparing Frequencies].

A confidence interval gives us a range of plausible values of the true odds ratio. In other words, a 95% confidence interval means we are 95% sure that the true odds ratio lies between these two values. This becomes important when the confidence interval is close to 0. If 0 is not in the range, then we are 95% sure that the two variables are dependent on each other. Also, if 0 is not in the confidence interval then we would get a p-value that is less than 0.05. Similarly, if 0 is in the 95% confidence interval then we are 95% sure that the two variables have no relationship with each other. This would correspond to a non-significant p-value, which has a value greater than 0.05. To understand how we interpret the odds ratio we will consider an odds ratio of 1.75 for a male. This indicates that a male is 1.75 times more likely to have a disease than a female [Zach, 2020].

4.5 Confusion Matrices and Related Terms

A confusion matrix is a table that summaries the actual categories and the predicted categories of a model [Rouse, 2018]. The columns represent the actual value, while the rows represent the predicted result. A true positive is when someone has a disease and is predicted to have a disease. A false positive is when someone does not have the disease, but they are predicted to have the disease. A true negative is when someone does not have a disease and they are predicted to not have it. Finally, a false negative is when someone does have the disease, but they are predicted to not have it [Mokobi, 2020]. Table 3 shows these and related terms such as sensitivity, specificity, accuracy, negative predictive value, and precision. It also shows the equations to calculate these. Sensitivity, also known as recall or true positive rate, is calculated using true positives and false negatives and tells us what percentage of patients with a disease were correctly identified [Starmer]. Ideally, we want the value of this to be 1. Specificity, also called true negative rate, is calculated using true negatives and false positives and tells us what percentage of patients without a disease were correctly identified [Starmer]. Like sensitivity, we want a value of 1. Accuracy uses all the rates and a value of 1 represents perfect accuracy. The negative predicted value is calculated using true negatives and false negatives. Finally, precision uses true positives and false positives, and like the others, a value of 1 is the best. If we want to correctly identify positive patients, we want our model to have a higher sensitivity, while if we want to correctly identify negatives, we should make sure our model has a higher specificity [Starmer].

Table 3 Confusion matrix with related terminology and calculations

		Actual		
		Positive	Negative	
Predicted	Positive	True Positive	False Positive	Precision= $\frac{TP}{TP+FP}$
	Negative	False Negative	True Negative	Negative Predicted Value= $\frac{TN}{TN+FN}$
		Sensitivity= Recall= True Positive Rate= $\frac{TP}{TP+FN}$	Specificity= True Negative Rate= $\frac{TN}{TN+FP}$	Accuracy= $\frac{TP+TN}{TP+TN\ +FN+FP}$

Another thing that uses true positive, false positive, false negative, and true negative values is the MCC equation that was discussed earlier. To calculate the MCC the following equation is used:

$$\text{MCC}= \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \qquad \textbf{(3)}$$

5. Data Subset Results

5.1. Correlations and Graphs

When using the subset of the kidney stone data, a lot of different correlations were looked at. These correlations were either the Pearson correlation or chi square test depending on the type of variable that the data was. Numerical variables had correlations done by the Pearson correlation. The chi square test was used on nominal variables. To be consistent, a p-value of less than 0.05 is significant. A correlation that has an absolute value less than 0.3 is considered a weak correlation. A correlation coefficient that has an absolute value between 0.3 and 0.5 is considered a moderate correlation. Any correlation coefficient whose absolute value is greater than 0.5 is said to represent a strong correlation [SPSS, 2020]. Switching the patient's diagnosis (icdtype) to a numerical value, 1 representing kidney stones (KS) and 0 representing abdominal and back pain (APB), it was possible to use the Pearson correlation test. After doing so, the correlation coefficient between icdtype and age was 0.108. This indicated that age and icdtype had a weak correlation with each other. However, the p-value that we got was 0.00075. This indicates that the correlation is significant. We conclude that age and icdtype have a significant, but weak correlation. Figure 1 shows a bar graph with a relation between age and icdtype. Looking at the blue bars, which indicate kidney stone patients, we see that people between 50 and 80 years old are more likely to have kidney stones than people who are younger. This supports the small positive correlation we found and suggests that age should be a factor that we consider when looking at the larger data set.

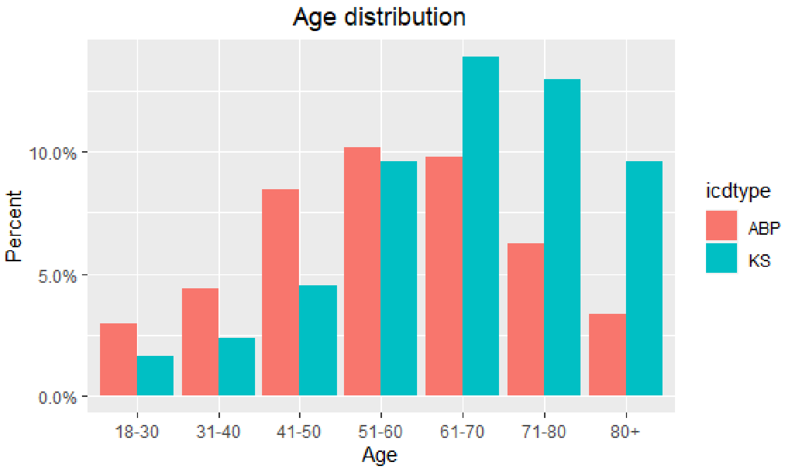


Figure 1 Bar graph showing icdtype based on patients age ranges

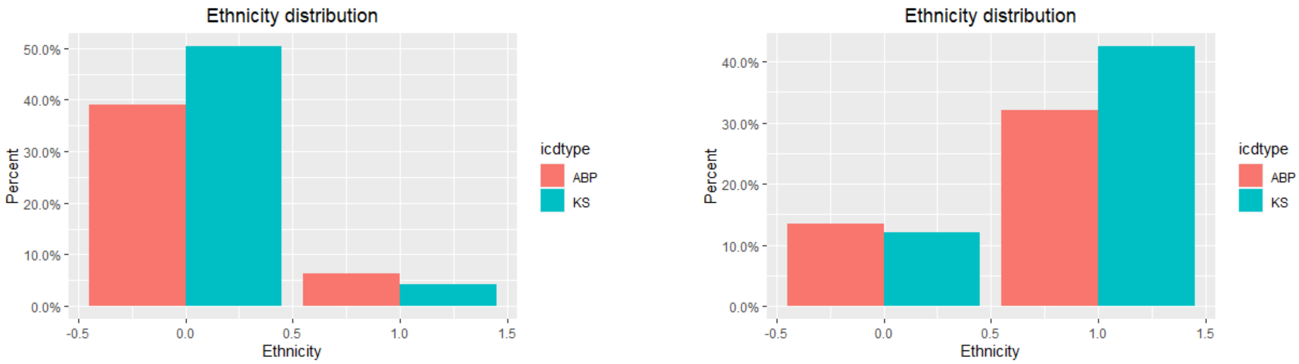


Figure 2 Bar graphs showing icdtype based on ethnicity. Left: black vs. non-black. Right: white vs. non-white.

Next, we looked to see if icdtype and gender were correlated. As both variables have nominal entries, the chi square test was used. The chi square test gave us a p-value of 0.1596. This means that our result was not significant, so we concluded that gender and icdtype were independent of each other. Similarly, we looked at icdtype and ethnicity. The chi square test for this resulted in a p-value of 0.0018, which is significant. So, from this we concluded that ethnicity and icdtype were dependent on each other. As there were multiple categories of ethnicities, we looked at two of the most populated ethnicities again with icdtype. This consisted of white and non-white, and black and non-black. Each done separately, we got corresponding chi square p-values of 0.002 and 0.0068, respectively. As both p-values were significant for each test, we can conclude that icdtype depends on white and non-white ethnicity, and icdtype depends on black and non-black ethnicity. The left graph of figure 2 shows the grouping of black and non-black, with black being represented by 1 and non-black being represented by 0. This graph shows that people who are black have a lower chance of getting a kidney stone than people who are non-black. The right graph of figure 2 shows the count of people with KS and ABP icdtypes, grouped by white and non-white. White is indicated by 1 and non-white is indicated by 0. From this graph we see that people who are white are more likely to have a kidney stone than people who are non-white. However, in both cases this data is very heavily populated by white patients and so the results we got are most likely biased. Doing these tests on a smaller set of data not only allowed for the learning of R, but it also gave us some key insight on factors to investigate when we look at the large data set. Based on the results from this subset of data, we see that age and ethnicity, specifically white and non-white and black and non-black could be important factors in determining a person's likelihood of forming kidney stones. The logistic regression results for this data subset can be found in appendix 2.

6. Kidney Stone Data Extracted From the MIMIC-III Critical Care Database

6.1 Data Cleaning/Introduction

After getting rid of all the duplicates we were able to get the number of data points down to 985, when looking at just KDS and ABP patients. Unfortunately, the urine results came with almost no information as all the slots were filled with NA. So, we could not utilize the urine information. However, with the other information we could calculate the estimated glomerular filtration rate or EGFR. To do this we used two different equations. We first used the MDRD Study Equation for Use with Standardized Serum Creatinine (Four-Variable Equation). We then calculated EGFR using the CKD-EPI Equation for Use with Standardized Serum Creatinine. Both equations are given in table 1. The reason why we calculated the EGFR using the CKD-EPI equation is because we had read a 2018 paper that had also used this equation. We wanted to compare our results to the results that they got [Chen, et.al., 2018].

6.2. Feature Selection

With so many different variables to choose from it was necessary for us to sift through them and find which ones are going to be most useful for our purposes. In our case we wanted to find the variables that would help us best sort people with certain symptoms into the groups of kidney stones and ABP, NCA, OKU, or OTH. Our initial step was to look at the correlation that each variable had with the icdtype. After running all the correlations, we found that none of them had very high correlations. EGFR, using the CKD-EPI equation, had the highest correlation with icdtype. This correlation had a value of -0.4103. The next highest correlations had absolute values that were just a little above |0.2|. More variables had absolute values in between |0.10| and |0.20|. Most of the variables had correlation values that landed below |0.10|. The first cut that we did was to not consider any of the variables that had a correlation value of less than |0.10|. We then ran logistic regressions using the variables that had correlation values higher than |0.10|. From these logistic regressions we were able to determine some of the most important variables for categorizing patients by icdtype. The variables that we ended up with were age, creatinine average, respiratory rate (resprate) mean, anion gap average, chloride average, hematocrit average, hemoglobin average, pH average, phosphate average, platelet average, blood urea nitrogen content (BUN) average, CO2 average, and red blood cell count (RBC) average. After narrowing it down to these variables we decided to run a regression with all these variables to see if they would all still be significant when categorizing icdtype. It turned out that age, creatinine average, respiratory rate (resprate) mean, chloride average, BUN average, and CO2 average were the only significant variables in this regression with p-values less than 0.05. We then took these variables and ran individual regressions. Also, since EGFR had the highest correlation value out of all the variables, we incorporated it into these regressions. A feature importance bar graph is shown in figure 3.

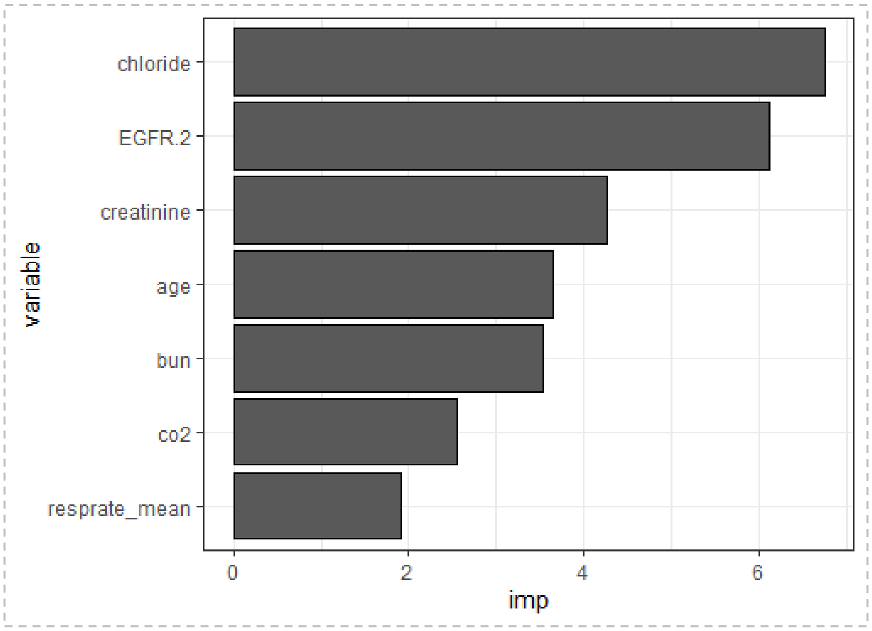


Figure 3 Feature importance plot

With a higher imp value representing a higher importance, we see that chloride seems to be the most important variable followed closely by EGFR. Creatinine, age, and BUN all have similar imp values at about 4. CO2 and mean respiratory rate (resprate_mean) are the two lowest variables in terms of their importance.

6.3 Logistic Regression Using Average Values

6.3.1 KDS VS ABP

We will go into depth for some of the regressions we did in this section, and then will only discuss the regression using the most correlated variables for each of the following sections. Summaries of all the regressions that we did will be discussed later. The first logistic regression that we will look at is the one that uses just age. Equation 4 gives the logistic regression equation using age. This set had 534 KDS patients and 451 ABP patients.

$$\log(\text{odds}) = -1.758 + 0.0312 \cdot \text{age} \qquad \textbf{(4)}$$

Figure 4 shows the logistic regression model and the corresponding probability-curve for equation 4. The index represents the patient's age. The value 1 indicates people who had kidney stones and the value 0 indicates people who had abdominal or back pain.

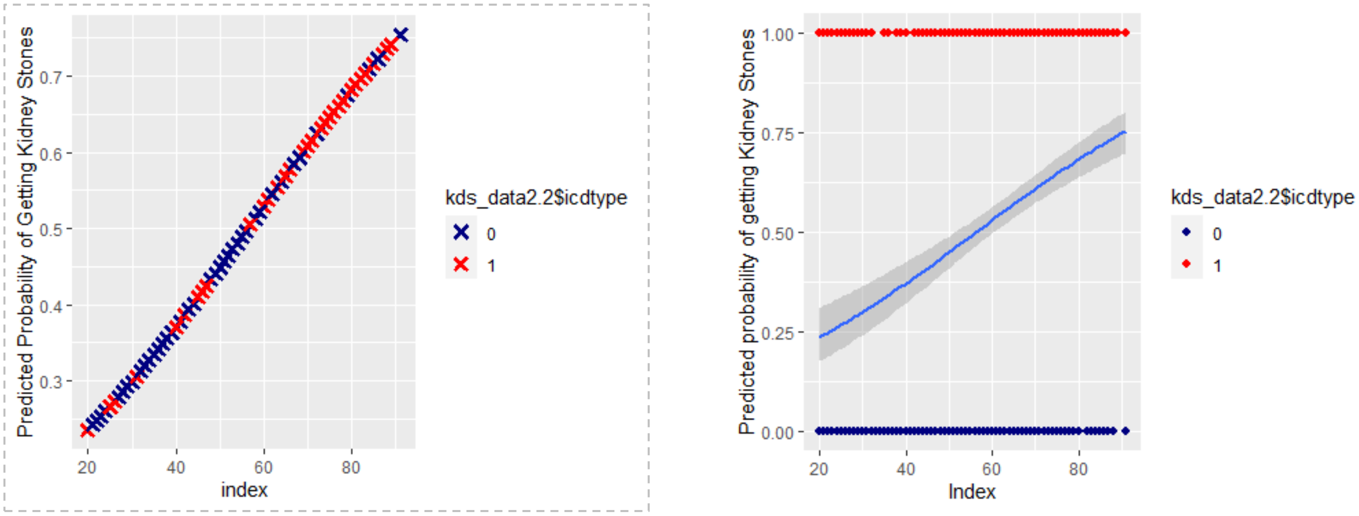


Figure 4 Logistic regression model (left) with corresponding probability-curve (right) for equation 4

Looking at the left graph of figure 4, we can see that there is a decent split at 50% predicted probability. Above 50% there are a lot of points that represent people with kidney stones and below that line most of the patients are diagnosed with abdominal or back pain. However, with the probability-curve we see that people of all ages are diagnosed with KDS or ABP. This indicates that this model may not be the best when categorizing patients. We also notice from both graphs in figure 4, that people from this data set seem to be starting with at least a 25% chance of developing a kidney stone. Even though this does not seem to be the best way to categorize, we wanted to know if there was a problem with this model. So, we created a binned residual plot, which is shown in figure 5. The grey lines represent the 95% confidence interval.

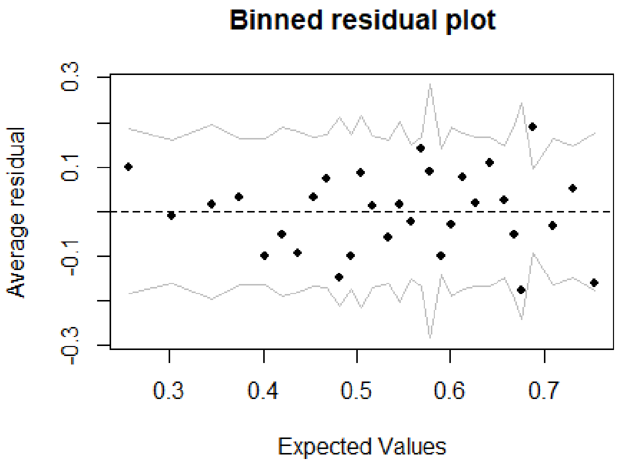


Figure 5 Binned residual plot using equation 4

Looking at the residual plot we do not see any major issues. It looks like we have an outlier in the data, but almost all the data points fall within the 95% confidence interval.

Because this model seems to be a reasonable model, we decided to look at the ROC curve and see what the AUC was for the curve. The ROC graph is shown in figure 6.

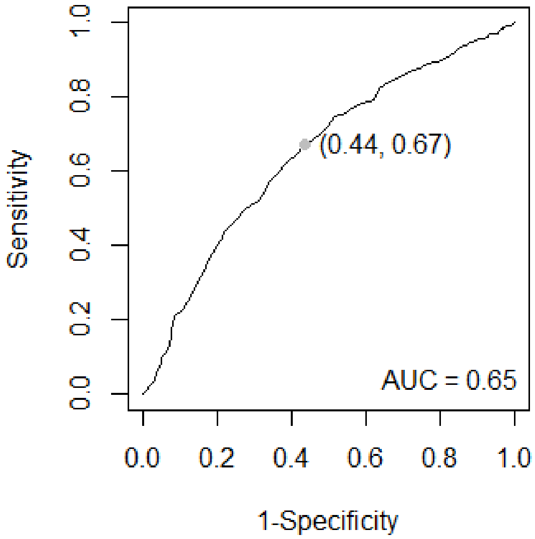


Figure 6 ROC graph using equation 4

This ROC graph had an area under the curve of 0.648, which is not a very good number. The closer to one the better. The sensitivity, specificity, and accuracy of this logistic regression is given in table 14. This AUC agrees with what we saw with the logistic model that age alone is not a very good way to categorize kidney stone patients or abdominal or back pain patients.

The next variable that we decided to look at was blood creatinine average. This variable was also significant in the heart disease data that can be found in appendix 3. It is interesting that this variable is important in both data sets. Although it makes sense it is used in both scenarios because ions present in the urine come from the blood plasma. Thus, the blood is an important factor in both heart disease and kidney stones. Equation 5 shows the logistic regression model that uses creatinine.

$$\log(\text{odds}) = -1.426+1.540*\text{creatinine\_avg} \qquad \textbf{(5)}$$

Similarly, for this model we made a logistic regression curve and a corresponding probability-curve. These two graphs are shown in figure 7. This logistic regression looks alright. One issue that we noticed immediately was that most of the points fall below an average creatinine level of 5. This consolidated all our data to the left side making it difficult to separate. However, looking at the s-curve, we can conclude that patients with the lowest creatinine levels seem more likely to not be diagnosed with kidney stones.

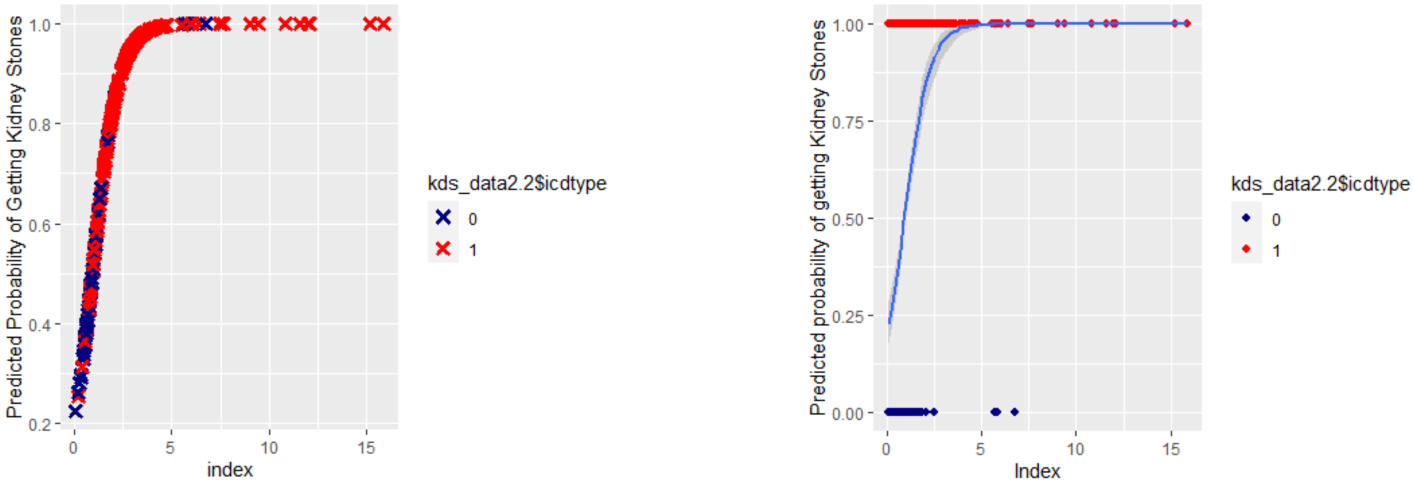


Figure 7 Logistic regression curve (left) and its corresponding probability-curve (right) for equation 5

While this variable seems to be a better categorizer on its own than age is, it doesn’t seem that this is the best way to do it. Either way we wanted to check and see if this model was a reasonable one to use, so we created a binned residual plot. This plot is shown in figure 8.

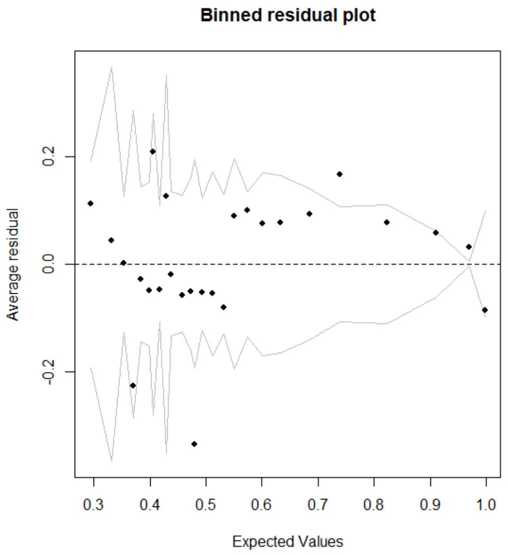


Figure 8 Residual plot for model using equation 5

Right away we notice some issues with the residual plot. Firstly, there are a good amount of points that lie outside of the 95% confidence interval. This tells us that there are more outliers in this data then we would normally expect. In addition, we notice at the beginning of the plot all the points fall below the 0 line and at the end of the plot they pretty much all fall above the 0 line. Points that fall below the line represent overpredicting and the points above the line would represent underpredicting. So, we can see that this model tends to overpredict at lower values of creatinine levels and underpredict at higher creatinine levels.

We decided to look at the ROC curve for this model and see what the area under the curve would be. Figure 9 shows the ROC curve.

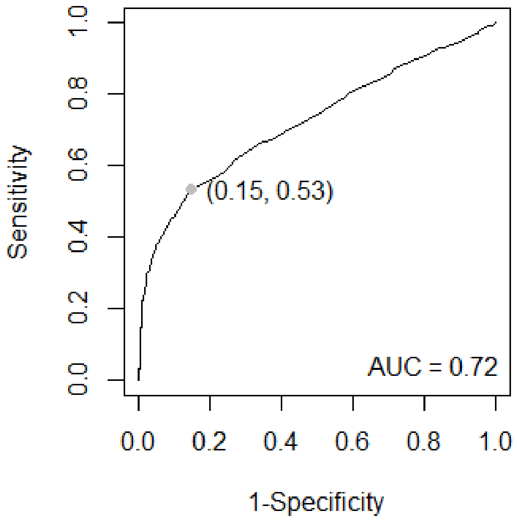


Figure 9 ROC curve for the model using equation 5

Despite having issues with overprediction and underprediction this ROC curve has a pretty good AUC. The AUC for this curve is 0.723. This is good and much higher than the one for equation 4. Based on the ROC curve it looks like we could do a decent job of maximizing the specificity for this data, but it would cost us the sensitivity.

The next variable that we looked at was the average blood urea nitrogen (BUN). In this case it is the average value. Equation 6 represents the logistic model using BUN.

log(odds)=-1.164 +0.068*bun_avg **(6)**

Figure 10 shows the logistic regression curve and the corresponding probability-curve for equation 6.

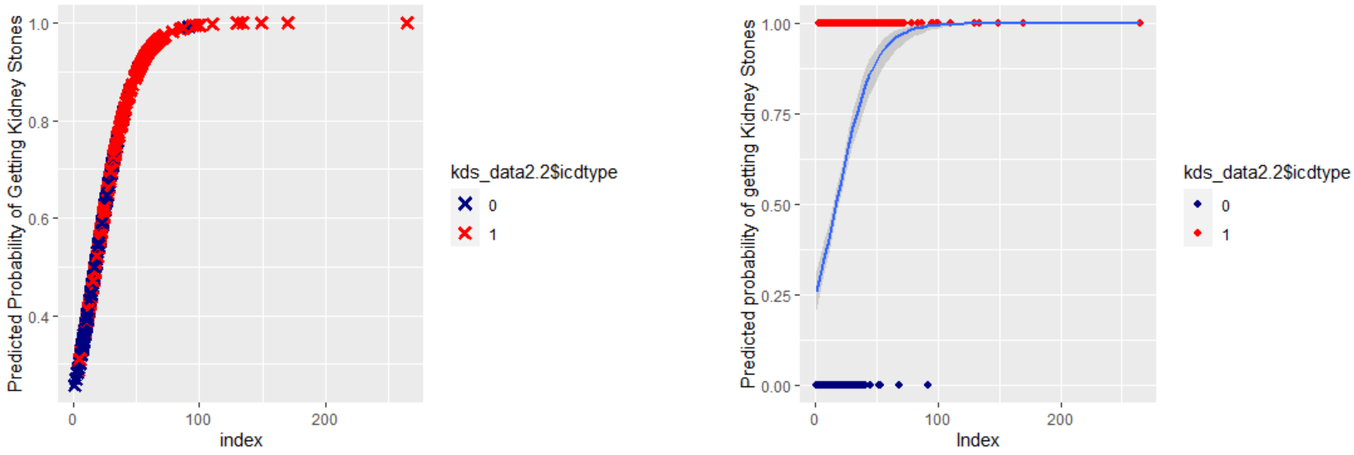


Figure 10 Logistic regression curve (left) and corresponding probability-curve (right) for equation 6

We see that most of the points fall below a BUN average of 100. This compiles most of the data points to a very small portion of our graph. In fact, nearly all the points that were not kidney stones fell below a BUN average of 50. In addition, we see that a lower BUN average has the lowest chance of being diagnosed with kidney stones. This variable seems to be just as good as creatinine and better than just using age in terms of categorizing patients.

We again wanted to look at the validity of this model, so we made a binned residual plot for equation 6. This binned residual plot is shown in figure 11.

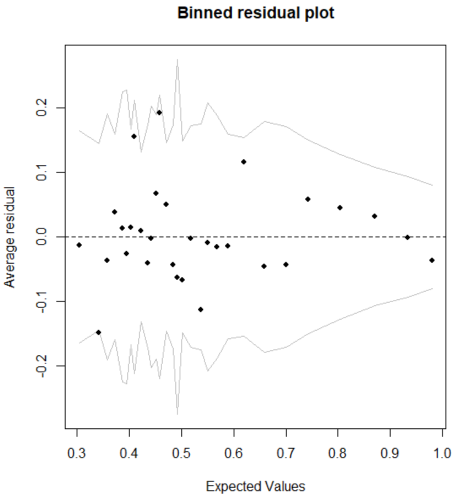


Figure 11 Binned residual plot for equation 6

Looking at the plot, these residuals look good. The points seem to be evenly distributed above and below the 0 line. In addition, there are only a couple of points that lie outside of the confidence interval. This indicates that there are not too many outliers. To compare just how similar bun average and creatine average were to each other, we looked at the ROC curve for equation 6. This plot is shown in figure 12.

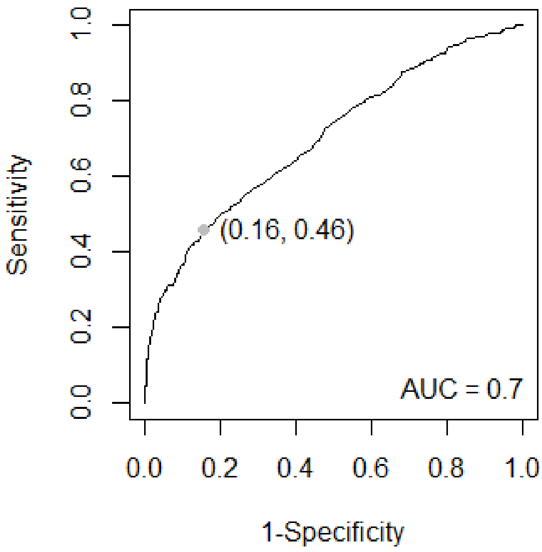


Figure 12 ROC curve for model using equation 6

This ROC curve had an area under the curve of 0.700. This value is slightly lower than the one we got using equation 6. This indicated to us that our suspicions about creatinine average and BUN average being similar in predicting kidney stones was correct.

The last individual regressions we looked at used EGFR. We did have two different equations that we used and wanted to compare them to each other. For simplicity we will call the EGFR calculated using the MDRD Study Equation, EGFR.1, and the EGFR calculated with the CKD-EPI equation, EGFR.2. Equation 7 shows the logistic regression equation using EGFR.1 and equation 8 shows the logistic regression equation using EGFR.2.

$$\log(odds)=1.590+ -0.017*EGFR.1 \qquad \textbf{(7)}$$

$$\log(odds)=4.202+ -0.0447*EGFR.2 \qquad \textbf{(8)}$$

Figure 13 shows the s-curves for equations 7 and 8. In both cases the index represents the EGFR values.

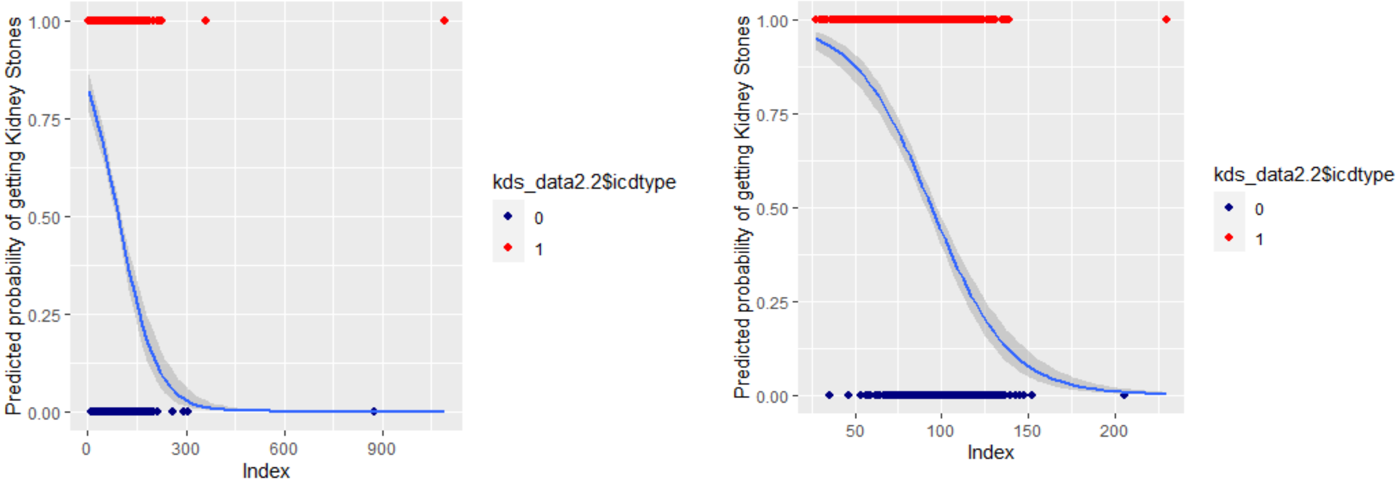


Figure 13 S-curve for equation 7 (left) and equation 8 (right)

These two s-curves look very similar to each other. Both seem to indicate that a higher EGFR results in a lower probability for kidney stones. The data seem to be more spread out in the one using equation 8 than the one using equation 7. As we mentioned earlier a higher EGFR is good in terms of kidney function. We see this idea a bit better in the s-curve given for equation 8, indicating that this might be the better EGFR calculation to use.

To look at the validity of these models we looked at the binned residual plots. Figure 14 shows these plots.

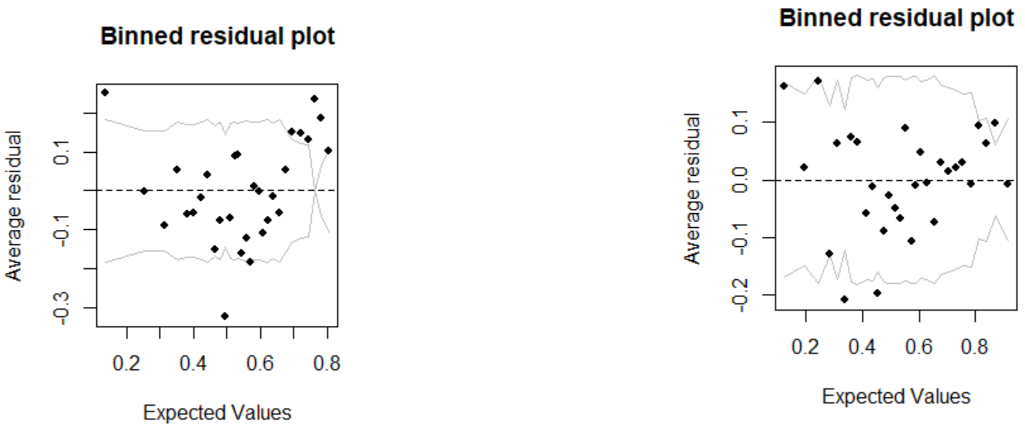


Figure 14 Binned residual plot for equation 7 (left) and equation 8 (right)

Overall, both plots indicate that the models we used were reasonable models. In both cases there seems to be a decent number of outliers in the data. This is indicated by the many points that fall outside the 95% confidence intervals.

To compare the logistic regressions of these two equations to each other, we looked at the ROC curve. Figure 15 shows these graphs.

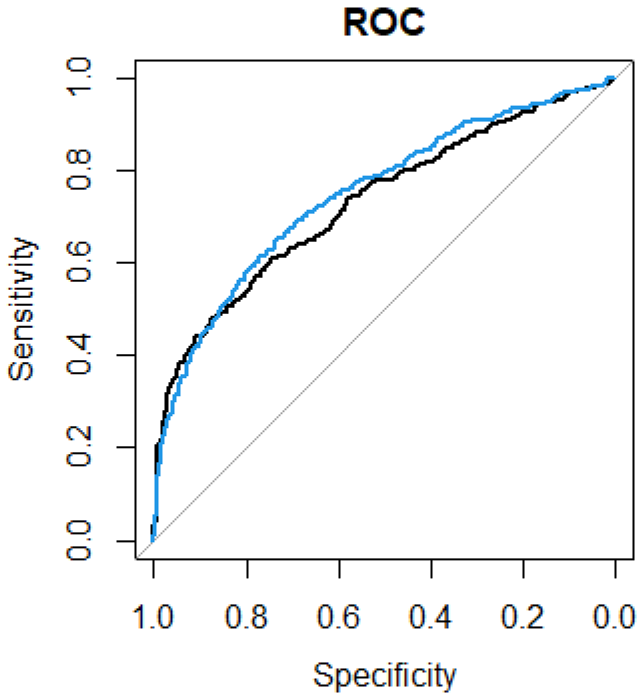


Figure 15 ROC curves for equation 7 (black line) and equation 8 (blue line)

The area under the curve for the ROC using equation 7, labeled by the black line, was 0.730. This had an optimum cut off point at (0.25, 0.61). The area under the curve for the ROC using equation 8, labeled by the blue line, was 0.748. This had an optimum cut off point at (0.26, 0.65). Both are good values for an AUC. We can see that EGFR.2 is slightly better at predicting than EGFR.1, but both are better at predicting than creatinine average, BUN average, and age.

The reason why we have equation 8 is so that we can compare the results to the ones that the 2018 paper got. It's hard to truly compare these values because we are looking at kidney stones vs abdominal and back pain, while the paper looked at patients diagnosed with genitourinary diseases, other conditions, and acute localized pain [Chen, et.al., 2018]. The diagnoses that we looked at would fall under other diseases. The 2018 paper had an AUC for their EGFR curve of .71 [Chen, et.al., 2018]. This is very close to what we got. Our value was slightly higher than theirs, but they had other diseases in that category that we did not look at. This was a very encouraging result as it looks like our results are what we should be getting from the data that we are using.

Finally, we looked at logistic regressions that used five different variables. The first one used chloride average, resprate mean, CO2 average, creatine average, and age. The second one used creatinine average, age, BUN average, CO2 average, and EGFR. For this EGFR we used the one calculated by the CKD-EPI equation. In both cases the s-curves did not show us much of anything as points were spread across both the top and the bottom. Because of this we did not include these graphs. The equation for the logistic regression using chloride average, resprate mean, CO2 average, creatine average, and age is given by equation 9.

log(odds)= -11.258 + 0.0864*chloride_avg+ 0.02871*age+ 0.852*creatinine_avg+ 0.0392*resprate_mean+ -0.0342*CO2_avg **(9)**

The logistic model using creatinine average, age, BUN average, CO2 average, and EGFR is given by equation 10.

log(odds)= 3.194 + 0.0211*bun_avg+ 0.0093*age+ 0.081*creatinine_avg+ -0.0225*EGFR.2+ -0.0748*CO2_avg **(10)**

Equation 10 is the regression model that used the most correlated variables for KDS VS ABP. To compare these two models, we looked at the ROC curves and the area under the curve. Figure 16 shows these ROC curves.

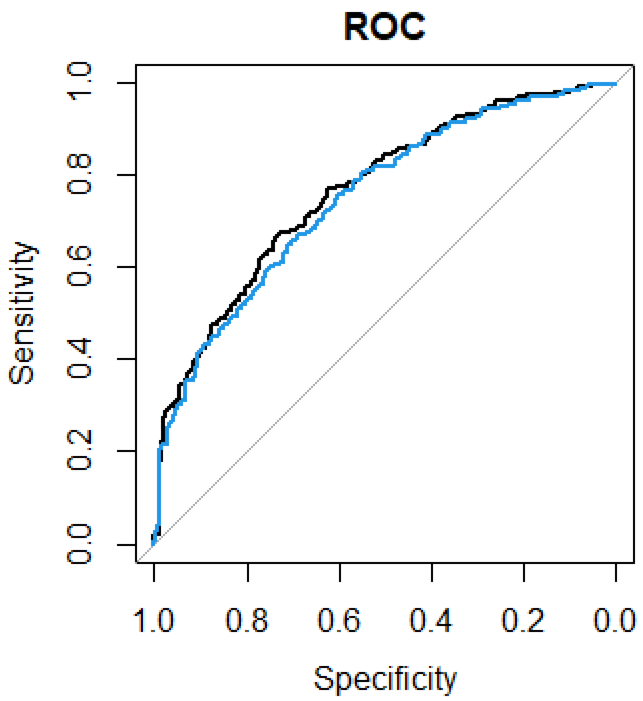


Figure 16 ROC curves for equation 9 (black line) and 10 (blue line)

The AUC for the ROC of equation 9, labeled by the black line, was 0.764. This had an optimum cut off point at (0.27, 0.67). The AUC for the ROC of equation 10, labeled by the blue line, was 0.753. This had an optimum cut off point at (0.24, 0.61). These are extremely close and indicate that we can predict just as well with each of these models. These AUC values are very close to the one we got just using the CKD-EPI equation for EGFR. From all this information it seems that the CKD-EPI EGFR predicts just as well as five variables do when categorizing KDS and ABP patients.

6.3.2 Random Forests for KDS VS ABP

We decided to look at the random forests produced for every equation that we just looked at. Some various decision trees can be found in appendix 4. We will first look at the random forest using equation 4. Table 4 shows the results of this random forest.

Table 4 Confusion matrix for random forest using equation 4

	KDS	ABP
KDS	370	147
ABP	233	202

From this random forest we got an OOB estimate of error rate of 39.92%. This tells us that 60.08% of the data was correctly identified by the random forest. This is not a great number as we want it to be as close to 100% as possible. The sensitivity for this random forest is 0.614 and the specificity is 0.579. Both values are not very good as we want these values as close to 1 as possible. It seems that using age alone for a random forest is not a very good categorizer, which agrees with our results from the logistic model.

Table 5 shows the results of the random forest that uses equation 5.

Table 5 Confusion matrix for random forest using equation 5

	KDS	ABP
KDS	271	244
ABP	105	328

From this random forest we got an OOB estimate of error rate of 36.81%. This tells us that 63.19% of the data was correctly identified by the random forest. This is slightly better than the random forest using equation 4, but not much better. The sensitivity for this random forest was 0.721 and the specificity was 0.573. The sensitivity for this random forest was much higher than the one we got when using just age, which means creatinine is better at predicting true positives than the model using just age. The specificities, however, are the same when using either age or creatinine. This random forest agrees with our initial conclusion that creatinine would be better used individually than age when categorizing kidney stone and abdominal and back pain patients.

Our next random forest was done using equation 6. Table 6 shows the results of this random forest.

Table 6 Confusion matrix for random forest using equation 6

	KDS	ABP
KDS	274	240
ABP	163	270

From this random forest we got an OOB estimate of error of 42.56%, which tells us 57.44% of the data was correctly categorized. This value is worse than both the one using age and creatinine. The sensitivity for this random forest was 0.627 and the specificity was 0.529. Both values are lower than the ones we got when just using creatinine. This result was a little surprising as the logistic regression indicated that BUN average and creatinine predict better than age. The result from this random forest suggests that BUN average should not be used individually to categorize patients when using a random forest.

Next, we will look at the two random forests done for the two EGFR formulas we used before. Table 7 will correspond to equation 7 which uses EGFR.1 or the MDRD Study Equation. Table 8 will correspond to equation 8 which uses EGFR.2 or the CKD-EPI equation.

Table 7 Confusion matrix for random forest using equation 7

	KDS	ABP
KDS	317	198
ABP	189	244

Table 8 Confusion matrix for random forest using equation 8

	KDS	ABP
KDS	326	189
ABP	188	245

The random forest from table 7 had an OOB estimate of error of 40.82%. This means 59.18% of the data was correctly identified. The sensitivity for table 7 was 0.626 and the specificity was 0.552. The random forest from table 8 had an OOB estimate of error of 39.77%, which means only 60.23% of the data was correctly identified. The sensitivity for table 8 was 0.634 and the specificity is 0.565. Both sensitivity and specificity were not good for each of these random forests. Neither one of these random forests performed better than the one that just used creatinine. When looking at individual variables for random forests it looks like it is best to use creatinine average.

We will now look at the results for the two random forests that each used five variables. Table 9 will correspond to the random forest using equation 9 and table 10 will correspond to the random forest using equation 10.

Table 9 Confusion matrix for random forest using equation 9

	KDS	ABP
KDS	233	81
ABP	97	127

Table 10 Confusion matrix for random forest using equation 10

	KDS	ABP
KDS	229	87
ABP	91	133

The random forest from table 9 had an OOB estimate of error of 33.09%. This means 66.91% of the data was correctly identified. The sensitivity for this random forest was 0.706 and the specificity was 0.611. The random forest from table 10 had an OOB estimate of error of 32.96%, which means only 67.04% of the data was correctly identified. The sensitivity for table 9 is 0.716 and the specificity is 0.605. These are the highest values that we have gotten for the random forests in accuracy, sensitivity, and specificity. This suggests that using more variables seems to categorize better than just 1 variable when using random forests. Overall, the logistic regressions seemed to predict better than the random forests.

6.3.3 KDS VS NCA

In this section and the following sections, we will discuss the logistic regression model that uses the most correlated variables. We will add summary tables at the end to look at all the regressions we did and compare them to each other. Running a correlation when looking at KDS and NCA we found that the most correlated variables were the Elixhauser comorbidity score, the 29th comorbidity group, which corresponds to drug abuse, the 30th comorbidity group, which corresponds to psychosis, and BUN. However, the 29th and 30th comorbidity groups were highly correlated with the Elixhauser score, so we decided to just use the Elixhauser score. Equation 11 shows the logistic equation for the most correlated variables.

$$\log(\text{odds})=-1.605+ (-0.0812) *\text{elixhauser\_vanwalraven}+ (-0.0300) *\text{bun} \qquad \textbf{(11)}$$

We then ran the regression and calculated an AUC, specificity, sensitivity, and accuracy for the ROC curve corresponding to this equation. The ROC curve is given in figure 17.

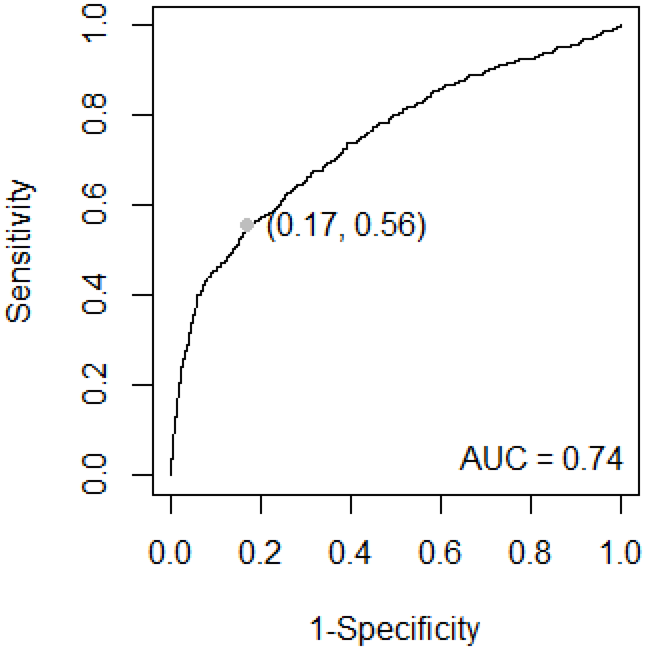


Figure 17 ROC curve for equation 11

The cutoff for the best values was at 0.17 on the x axis and 0.56 on the y-axis. The AUC for this regression had a value of 0.744, which is not fantastic, but it is not bad either. This graph had a specificity of 0.829 and a sensitivity of 0.557. This is a pretty good specificity value, but the sensitivity value is not very good. Finally, we got an accuracy of 0.821, which is a pretty good value. Overall, this regression does a pretty good job distinguishing between groups as it has an accuracy of 82%. In addition to the most correlated variables we looked at regressions using age, BUN, creatinine, EGFR, Elixhauser score, and two regressions using five variables. These regressions will be discussed in section 6.6.

6.3.4 KDS VS OKU

In this section we will look at the regression of the most correlated variables for KDS VS OKU. Running a correlation for all the variables we found that the most correlated variables were BUN, bands, and creatinine. Equation 12 shows the logistic regression for the most correlated variables.

log(odds)= -2.443+(-0.000097) *bun +0.0348*bands+0.470*creatinine **(12)**

We looked at the ROC curve for this regression model and calculated the AUC, specificity, sensitivity, and accuracy. The ROC curve is shown in figure 18.

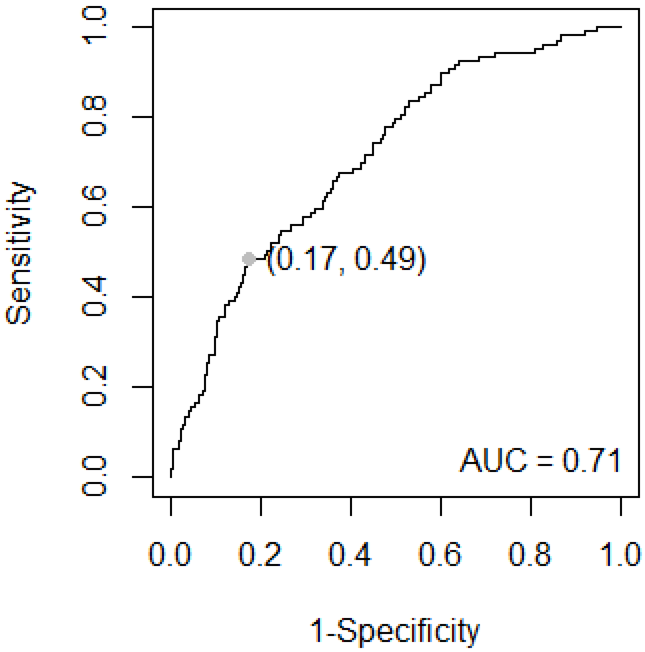


Figure 18 The ROC curve for equation 12

The cutoff value for the best numbers was 0.17 on the x-axis and 0.49 on the y-axis. This curve had an AUC of 0.713, which is pretty good, but not great. The specificity for this graph was 0.826 and had a sensitivity of 0.486. The accuracy of this ROC curve had a value of 0.763. Comparing this to the regression for KDS VS NCA we see that the most correlated variables did a better job categorizing for KDS VS NCA than the most correlated variables did for KDS VS OKU. Regressions were done for age, BUN, creatinine, EGFR, Elixhauser score, and two regressions using five variables which are summarized in section 6.6.

6.3.5 KDS VS OTH

In this section we look at KDS VS OTH. Again, we ran a correlation for all the variables, and we found that BUN and creatinine were the most correlated variables for KDS VS OTH. Equation 13 shows the logistic regression equation for the most correlated variables.

$$\log(odds)=-4.684+ 0.0318*BUN+ 0.1838*creatinine \quad \textbf{(13)}$$

We looked at the ROC curve for this equation and calculated the AUC, specificity, sensitivity, and accuracy. Figure 19 shows the ROC graph for this equation.

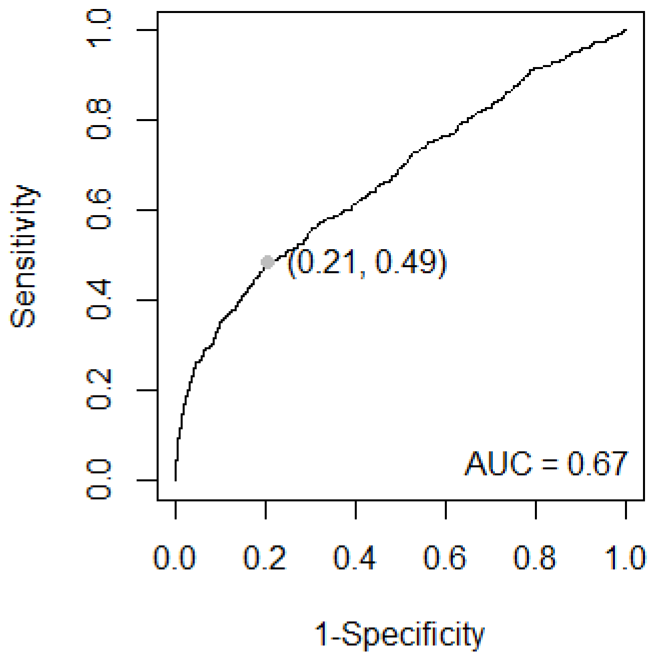


Figure 19 The ROC curve for equation 13

The cutoff for this ROC curve that gave the best values was 0.21 on the x-axis and 0.49 on the y-axis. The AUC of this graph had a value of 0.671. The specificity was 0.794 and the sensitivity was .487. The accuracy for this regression was 0.787. Looking at all four of the regressions we see that the most correlated variables best categorized KDS and NCA.

6.4 Logistic Regression for First Lab Results

6.4.1 KDS VS ABP

We wanted to compare the logistic regressions that were created when we used the average lab results and the first lab result values. This set had 534 KDS patients and 451 ABP patients. We are looking at the most correlated variables, which were creatinine, age, bun, CO2, and EGFR. The values were slightly changed for the logistic regression. Equation 14 shows the logistic regression created with the most correlated variables.

log(odds)= 2.0366+ 0.0160*age+ 0.0196*bun+ 0.1067*creatinine+ (-0.0166) *EGFR.2+ (-0.0664)*CO2 **(14)**

The AUC for this regression had a value of 0.744. The specificity was 0.670 and the sensitivity was 0.713. The accuracy was 0.695. Comparing this to the average lab result values we see that we get slightly lower values for both AUC and specificity, while the sensitivity and accuracy slightly increase. Based on the accuracy that we get it seems that the first lab results do slightly better at categorizing when using the same variables than the average lab values do.

6.4.2 KDS VS NCA

Looking at the first lab results for KDS VS NCA we again looked at the most correlated variables, which are the same as before, and compared the two logistic regressions that we got. This data had 534 KDS patients and 16,701 NCA patients. Equation 15 shows the logistic equation that we get using the first lab results.

$$\log(\text{odds})=-1.639+ (-0.0835) *\text{elixhauser\_vanwalraven}+ (-0.0275) *\text{bun} \tag{15}$$

The AUC for the corresponding ROC curve had a value of 0.742. The specificity was 0.835 and the sensitivity was 0.540. The accuracy of this regression was 0.825. Comparing this to the values we got when we used the average lab results, we see that the AUC's were about the same, while specificity and accuracy slightly increased, and sensitivity slightly declined. Looking at the overall accuracy of both models we conclude that the average lab results were best in categorizing KDS and NCA patients.

6.4.3 KDS VS OKU

We also looked at the first lab results for KDS VS OKU. Using the same most correlated variables as before, we did a logistic regression and formed equation 16. This data had 534 KDS patients and 4,620 OKU patients.

$$\log(\text{odds})=2.4381+0.0007*\text{bun} +0.0375*\text{bands}+0.4378*\text{creatinine} \tag{16}$$

The AUC for the ROC curve was 0.700. The specificity was 0.507 and the sensitivity was 0.789. The accuracy for this logistic regression was 0.560. Comparing this to the average lab values we see that the AUC slightly decreased. The specificity decreased significantly, and the sensitivity increased significantly. The accuracy took a big hit, decreasing by a lot. We conclude, based on the accuracy, that the average lab results do a better job at categorizing these two groups.

6.4.4 KDS VS OTH

Again, we looked at the first lab result values for KDS VS OTH. Using the same variables as before, we got the logistic regression equation, given by equation 17. This data had 534 KDS patients and 24,295 OTH patients.

$$\log(\text{odds})= -4.638+ 0.0283*\text{bun}+ 0.1836*\text{creatinine} \tag{17}$$

The AUC value for the corresponding ROC graph was 0.669. The specificity was 0.782 and the sensitivity was 0.483. The overall accuracy of this regression equation was 0.775. Comparing these values to the regression using the average lab results, we see that all the values were very similar but had slightly smaller numbers when using the first lab results. This suggests that the average lab results, result in a slightly better prediction.

6.5 First Admittance and First Lab Result Logistic Regression

6.5.1 KDS VS ABP

Finally, we decided to compare the previous regressions to the regression created using the first lab values that were obtained on the patients' very first admittance to the hospital. We again looked at the most correlated variables which in this case were EGFR, CO2, BUN, age, and creatinine. This set of data included 332 ABP patients and 389 KDS patients. Equation 18 gives the logistic regression equation.

$$\log(odds)= 2.646+ 0.0137*age+ 0.0096*bun+ 0.0541*creatinine+ (-0.0127) *EGFR.2+ (-0.0558) *CO2 \quad \textbf{(18)}$$

This logistic regression had an AUC of 0.718 for the ROC graph. The specificity had a value of 0.591 and sensitivity of 0.746. The overall accuracy of this regression was 0.683. Looking at the overall accuracy of the three different regressions, we see that using the first lab results produced the highest accuracy. The next highest accuracy came when using the first lab results and the patients' first admittance. The worst performing regression came when using the average values.

6.5.2 KDS VS NCA

Similarly, we looked at the first lab results and the first admission to compare its regression to the other two regressions we have looked at for KDS VS NCA. This data had 389 KDS patients and 9,845 NCA patients. Again, we used the most correlated variables, which were the Elixhauser comorbidity score and BUN. Equation 19 gives the logistic regression equation.

$$\log(odds)=-1.417+ (-0.0924) *elixhauser\_vanwalraven+ (-0.0033) *bun \quad \textbf{(19)}$$

The AUC for the corresponding ROC graph had a value of 0.754. The specificity was 0.898 and the sensitivity was 0.508. The overall accuracy of this regression was 0.883. Comparing the accuracies to the other two regressions, we see that this model does better than the one using average values and the one using just the first lab results. It has a larger accuracy by about 0.06. The second highest result came when using the first lab result values, while the lowest result occurred when using the average values.

6.5.3 KDS VS OKU

The most correlated variables when looking at KDS VS OKU were BUN, bands, and creatinine. We again got the first lab results from the first admittance and ran a regression. This data had 389 KDS patients and 3,249 OKU patients. Equation 20 shows the logistic regression equation.

$$\log(odds)=-2.704+(-0.0004) *bun +0.0307*bands+0.7815*creatinine \quad \textbf{(20)}$$

The AUC for the corresponding ROC graph had a value of 0.722. The specificity was 0.793 and the sensitivity was 0.588. The overall accuracy for this model was 0.751. Comparing this result to the other two results that we got, we see that this regression had the highest accuracy. The second highest result came when we used the average lab results. A significantly lower accuracy occurred when we used only the first lab results.

6.5.4 KDS VS OTH

Just like we have done previously we looked at the first lab results and the first admittance for KDS VS OTH. This data had 389 KDS patients and 24,292 OTH patients. The most correlated variables were still BUN and creatinine. Equation 21 represents the logistic regression equation.

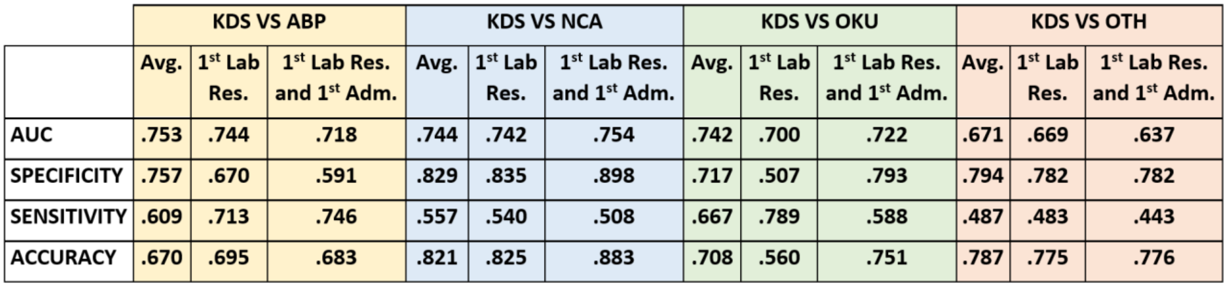
$$\log(odds)= -4.667+ 0.0266*bun+ 0.1509*creatinine \qquad \textbf{(21)}$$

The AUC for the corresponding ROC graph was 0.637. The specificity was 0.782 and the sensitivity was 0.443. The overall accuracy of this regression was 0.776. Now, let's compare this result to the other two results we got. The accuracy we got for this result was nearly identical to the accuracy we got when using just the first lab result. However, both results are slightly lower than the results that we got when we used the average values. We will compare other regression results using the three different data values in the next section.

6.6 Logistic Regression and Statistical Summaries

To make comparisons between the different values of data that we looked at, we made tables for each individual regression we did. Table 11 shows the logistic regressions using the most correlated variables, which have been discussed in detail in sections 6.3, 6.4, and 6.5.

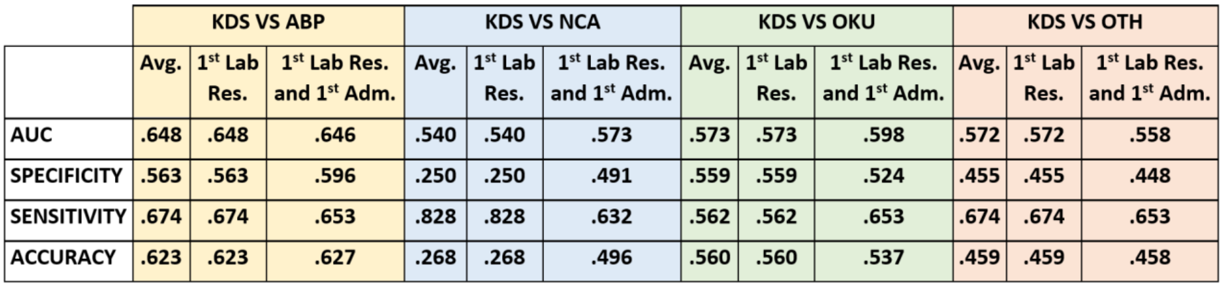
Table 11 Logistic regression using most correlated variables summary



	KDS VS ABP			KDS VS NCA			KDS VS OKU			KDS VS OTH		
	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.
AUC	.753	.744	.718	.744	.742	.754	.742	.700	.722	.671	.669	.637
SPECIFICITY	.757	.670	.591	.829	.835	.898	.717	.507	.793	.794	.782	.782
SENSITIVITY	.609	.713	.746	.557	.540	.508	.667	.789	.588	.487	.483	.443
ACCURACY	.670	.695	.683	.821	.825	.883	.708	.560	.751	.787	.775	.776

The next regression that we looked at used just age. Table 12 shows the results of all the logistic regressions that we ran using just age.

Table 12 Logistic regression using age summary

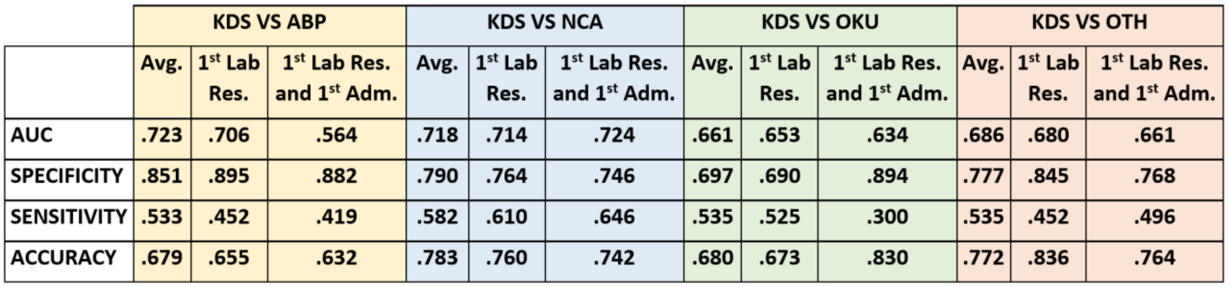


	KDS VS ABP			KDS VS NCA			KDS VS OKU			KDS VS OTH		
	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.
AUC	.648	.648	.646	.540	.540	.573	.573	.573	.598	.572	.572	.558
SPECIFICITY	.563	.563	.596	.250	.250	.491	.559	.559	.524	.455	.455	.448
SENSITIVITY	.674	.674	.653	.828	.828	.632	.562	.562	.653	.674	.674	.653
ACCURACY	.623	.623	.627	.268	.268	.496	.560	.560	.537	.459	.459	.458

Looking at table 11 we see that we got the exact same results when we used the average lab values and the first lab values for each category. We also notice that the accuracy was the highest for KDS VS ABP and KDS VS NCA when using the first lab results from the first admission, while it was highest for KDS VS OKU and KDS VS OTH, when using the average lab values. Regardless of what category of patients we were looking at, we see that none of the accuracies were very good, which indicates that age alone is not the best categorizer.

Next, we decided to look at creatinine as the only variable and run logistic regressions for each scenario. Table 13 shows the summary results of the regressions when just using creatinine.

Table 13 Logistic regression using just creatinine summary

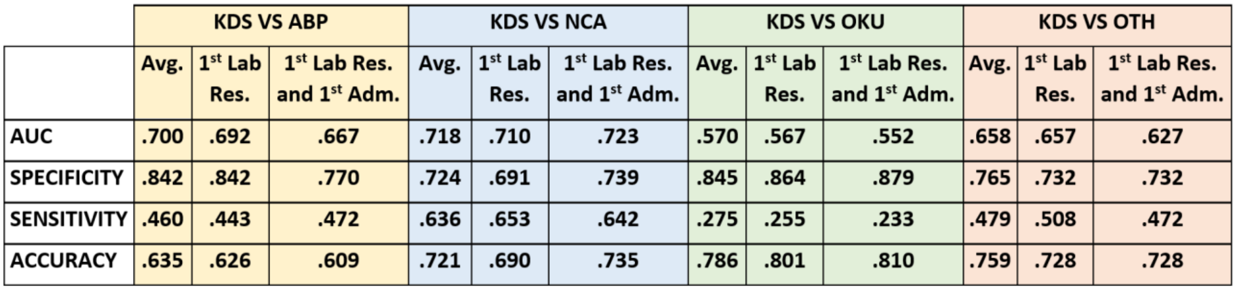


	KDS VS ABP			KDS VS NCA			KDS VS OKU			KDS VS OTH		
	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.
AUC	.723	.706	.564	.718	.714	.724	.661	.653	.634	.686	.680	.661
SPECIFICITY	.851	.895	.882	.790	.764	.746	.697	.690	.894	.777	.845	.768
SENSITIVITY	.533	.452	.419	.582	.610	.646	.535	.525	.300	.535	.452	.496
ACCURACY	.679	.655	.632	.783	.760	.742	.680	.673	.830	.772	.836	.764

Looking at table 13, we see that the highest accuracy results for KDS VS ABP and KDS VS NCA came when we used the average lab values. For KDS VS OKU the highest accuracy came when using the first lab results from the first admission. For KDS VS OTH the highest result came when looking at just the first lab results. These accuracy values are much higher than the ones we got when we were just looking at age alone. This indicated to us that using creatinine is better than using age if we wanted to categorize using only one variable.

Our next regression that we looked at was one that used only the BUN values. Table 14 shows the summary of these logistic regressions.

Table 14 Logistic regression using only BUN values summary

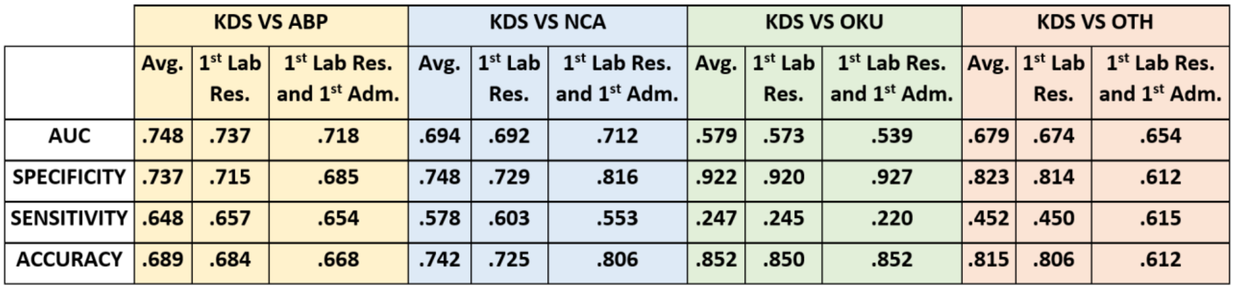


	KDS VS ABP			KDS VS NCA			KDS VS OKU			KDS VS OTH		
	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.
AUC	.700	.692	.667	.718	.710	.723	.570	.567	.552	.658	.657	.627
SPECIFICITY	.842	.842	.770	.724	.691	.739	.845	.864	.879	.765	.732	.732
SENSITIVITY	.460	.443	.472	.636	.653	.642	.275	.255	.233	.479	.508	.472
ACCURACY	.635	.626	.609	.721	.690	.735	.786	.801	.810	.759	.728	.728

We again see a split between which values produce the best results. KDS VS ABP and KDS VS OTH get the best results when the average lab values are used. The other two groups get their best results when looking at the first lab values from the first admission. Comparing this to the creatinine regressions we see that we found slightly higher values of accuracy when using just creatinine than we did in their corresponding categories using just BUN values.

Our next individual regression that we looked at was the EGFR values. We calculated these EGFR values using the CKD-EPI Equation for Use with Standardized Serum Creatinine. Table 15 gives the summaries of these logistic regressions.

Table 15 Logistic regression using just EGFR values summary

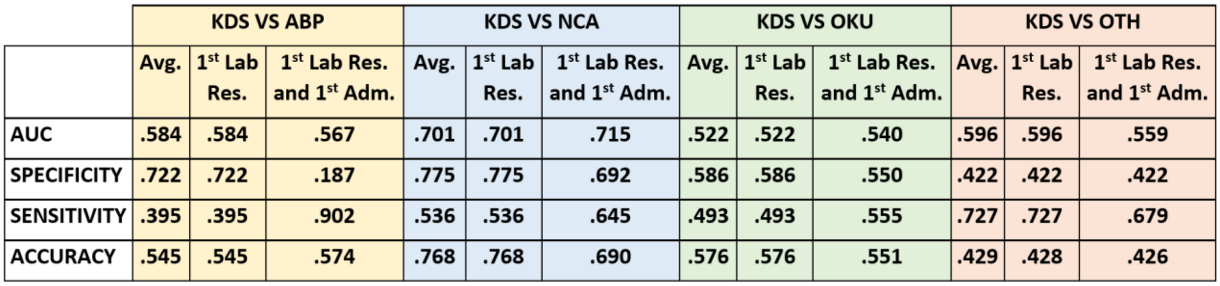


	KDS VS ABP			KDS VS NCA			KDS VS OKU			KDS VS OTH		
	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.
AUC	.748	.737	.718	.694	.692	.712	.579	.573	.539	.679	.674	.654
SPECIFICITY	.737	.715	.685	.748	.729	.816	.922	.920	.927	.823	.814	.612
SENSITIVITY	.648	.657	.654	.578	.603	.553	.247	.245	.220	.452	.450	.615
ACCURACY	.689	.684	.668	.742	.725	.806	.852	.850	.852	.815	.806	.612

The best accuracy for KDS VS ABP, KDS VS OKU, and KDS VS OTH all came when using the average lab results. We do note that the accuracy for KDS VS OKU was the same for both the average lab results and the first lab results from the 1st admission. KDS VS NCA got its best result when using the first lab values from the first admission. In all categories except KDS VS OTH the highest accuracies from just using EGFR were higher than the highest accuracies when just using creatinine. The accuracy using creatinine was .021 higher than that using just EGFR for KDS VS OTH. From this we concluded that using EGFR was the better variable to use if we were just looking at a single variable to categorize patients.

Our final individual variable that we looked at was the Elixhauser score. Because this score looks at one's likelihood of dying in a hospital, we thought it would be interesting to see how well its logistic regression would do when categorizing patients. Table 16 shows the summaries for these logistic regressions.

Table 16 Logistic regression summaries when using only Elixhauser score



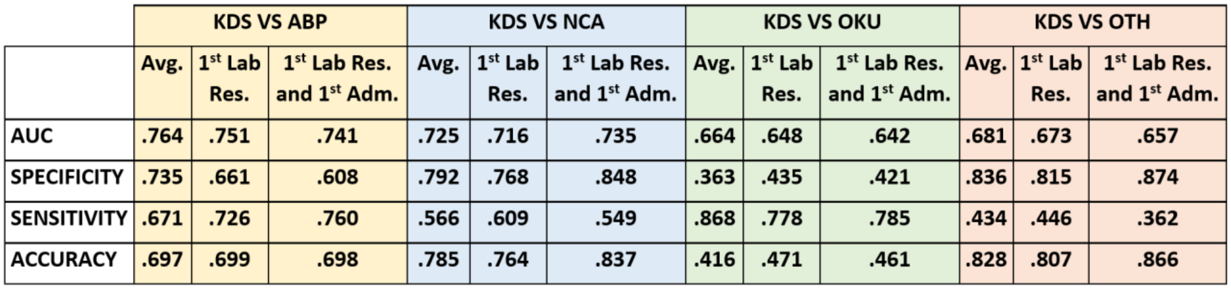
	KDS VS ABP			KDS VS NCA			KDS VS OKU			KDS VS OTH		
	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.
AUC	.584	.584	.567	.701	.701	.715	.522	.522	.540	.596	.596	.559
SPECIFICITY	.722	.722	.187	.775	.775	.692	.586	.586	.550	.422	.422	.422
SENSITIVITY	.395	.395	.902	.536	.536	.645	.493	.493	.555	.727	.727	.679
ACCURACY	.545	.545	.574	.768	.768	.690	.576	.576	.551	.429	.428	.426

We notice that the average lab results and the first lab results produced the same values for each of the four groupings. We see that other than KDS VS ABP, which got its best accuracy when looking at the first lab results from the first admission, the best accuracies came when using the average lab results or the 1st lab results, since they produced the same results. It is also clear that these regressions had very low accuracies except for KDS VS NCA, which had a decent accuracy. We assume that Elixhauser scores perform better for KDS VS NCA patients because the diseases that are categorized as NCA are more likely to cause patient death, while in the hospital. Since the models did poorly overall, we concluded that using the Elixhauser score was not a very good way to categorize these patients. So, we can conclude that, when used individually, the EGFR values gave us the best results.

Now we will investigate two regressions that each used 5 variables. Our first regression is one that does not use the EGFR values. The five variables that it uses are resprate mean, creatinine, CO2, age, and chloride. Table 17 shows the summaries of these regressions.

Table 17

Regression summaries when using resprate mean, creatinine, age, chloride, and CO2

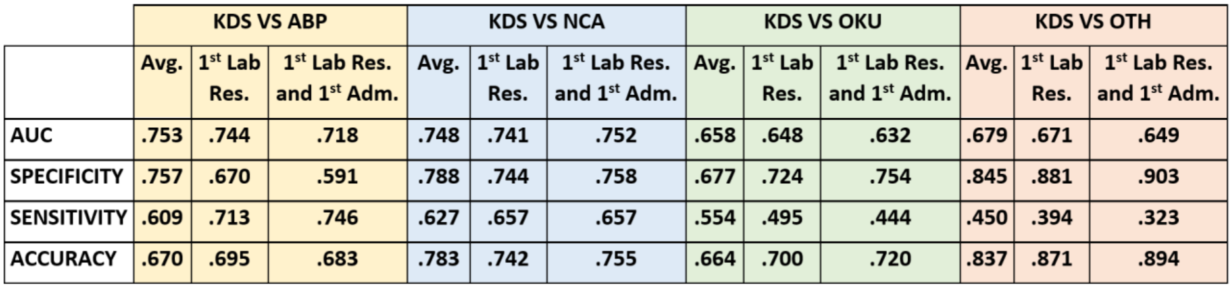


	KDS VS ABP			KDS VS NCA			KDS VS OKU			KDS VS OTH		
	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.
AUC	.764	.751	.741	.725	.716	.735	.664	.648	.642	.681	.673	.657
SPECIFICITY	.735	.661	.608	.792	.768	.848	.363	.435	.421	.836	.815	.874
SENSITIVITY	.671	.726	.760	.566	.609	.549	.868	.778	.785	.434	.446	.362
ACCURACY	.697	.699	.698	.785	.764	.837	.416	.471	.461	.828	.807	.866

Looking at table 17 we see that the accuracy is very similar for KDS VS ABP no matter what data you use. Its highest value does come when the 1st lab results are used. KDS VS OKU also has its highest accuracy value when the first lab results were used. Both KDS VS NCA and KDS VS OTH have their highest accuracy values when the first lab results from the first admission are used. Overall, this regression seems to do well when looking at OTH and NCA patients and does more poorly when looking at ABP and OKU patients.

Finally, we looked at a regression that used five variables and included EGFR. The other four variables that were used were age, creatinine, BUN, and CO2. Table 18 summarizes the results that we got from these regressions.

Table 18 Regression summaries when using EGFR, age, creatinine, CO2, and BUN



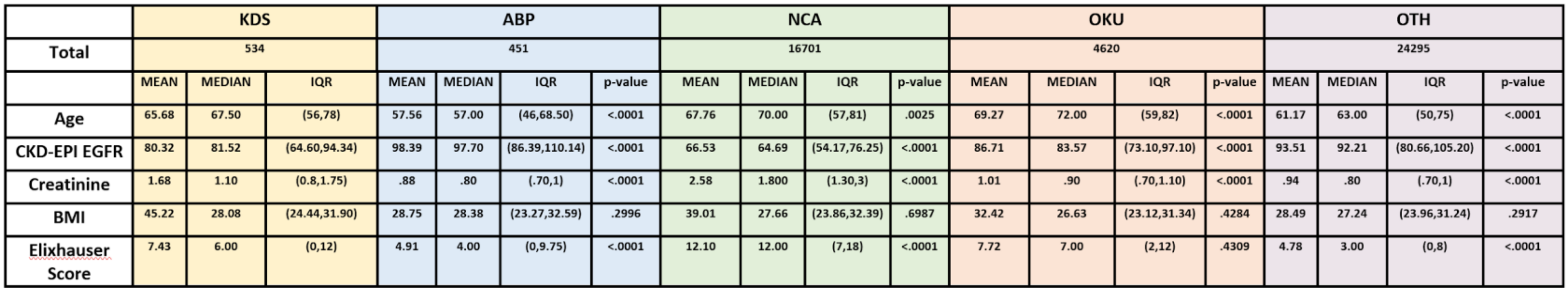
	KDS VS ABP			KDS VS NCA			KDS VS OKU			KDS VS OTH		
	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.	Avg.	1st Lab Res.	1st Lab Res. and 1st Adm.
AUC	.753	.744	.718	.748	.741	.752	.658	.648	.632	.679	.671	.649
SPECIFICITY	.757	.670	.591	.788	.744	.758	.677	.724	.754	.845	.881	.903
SENSITIVITY	.609	.713	.746	.627	.657	.657	.554	.495	.444	.450	.394	.323
ACCURACY	.670	.695	.683	.783	.742	.755	.664	.700	.720	.837	.871	.894

KDS VS ABP gets its best results when the first lab results are used. KDS VS NCA on the other hand gets its best results when the average results are used. Both KDS VS OKU and KDS VS OTH get their best results when the first lab results from the first admission are used. Immediately we noticed improvement in the KDS VS OKU results when compared to the other five variable regression. The other three groups stayed relatively the same. Because of these more consistent values, with them all being over 0.600, we believe that it is better to use this last regression when categorizing these patients.

Overall, we seem to get a split decision on which set of data produces the best results. In some cases, we found that the average lab results produced better numbers and in other cases found that the first lab results from the first admission produced better numbers. Depending on what variables one would want to use and what group one is looking at, it is possible to find the best data to use. However, there is not a data set that we can definitively say will always give the best outcomes or results.

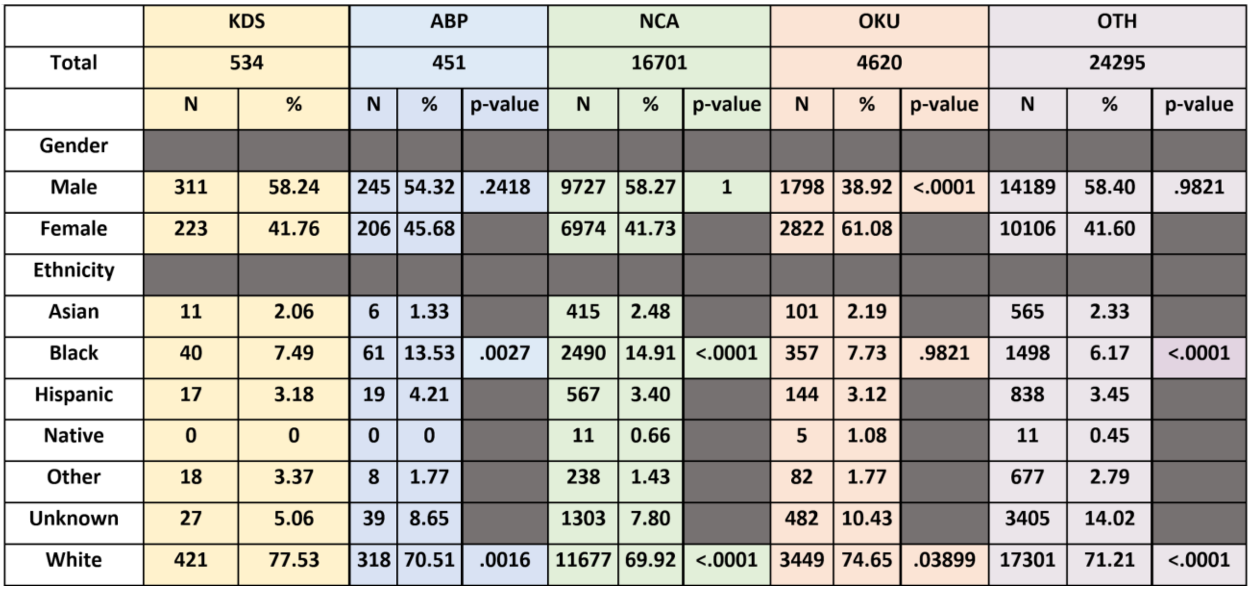
In addition to these logistic summaries we did some quick statistical summaries of the variables we used the most. We also looked at how the ethnicities and genders fanned out over the different groups of patients. Table 19 shows the statistical summaries of some of the variables we used, while table 20 shows the groupings of gender and ethnicity by their icdtype.

Table 19 Statistical summaries of commonly used variables



	KDS			ABP				NCA				OKU				OTH			
Total	534			451				16701				4620				24295			
	MEAN	MEDIAN	IQR	MEAN	MEDIAN	IQR	p-value	MEAN	MEDIAN	IQR	p-value	MEAN	MEDIAN	IQR	p-value	MEAN	MEDIAN	IQR	p-value
Age	65.68	67.50	(56,78)	57.56	57.00	(46,68.50)	<.0001	67.76	70.00	(57,81)	.0025	69.27	72.00	(59,82)	<.0001	61.17	63.00	(50,75)	<.0001
CKD-EPI EGFR	80.32	81.52	(64.60,94.34)	98.39	97.70	(86.39,110.14)	<.0001	66.53	64.69	(54.17,76.25)	<.0001	86.71	83.57	(73.10,97.10)	<.0001	93.51	92.21	(80.66,105.20)	<.0001
Creatinine	1.68	1.10	(0.8,1.75)	.88	.80	(.70,1)	<.0001	2.58	1.800	(1.30,3)	<.0001	1.01	.90	(.70,1.10)	<.0001	.94	.80	(.70,1)	<.0001
BMI	45.22	28.08	(24.44,31.90)	28.75	28.38	(23.27,32.59)	.2996	39.01	27.66	(23.86,32.39)	.6987	32.42	26.63	(23.12,31.34)	.4284	28.49	27.24	(23.96,31.24)	.2917
Elixhauser Score	7.43	6.00	(0,12)	4.91	4.00	(0,9.75)	<.0001	12.10	12.00	(7,18)	<.0001	7.72	7.00	(2,12)	.4309	4.78	3.00	(0,8)	<.0001

Table 20 Groupings of ethnicity and gender by icdtype



	KDS		ABP			NCA			OKU			OTH		
Total	534		451			16701			4620			24295		
	N	%	N	%	p-value	N	%	p-value	N	%	p-value	N	%	p-value
Gender														
Male	311	58.24	245	54.32	.2418	9727	58.27	1	1798	38.92	<.0001	14189	58.40	.9821
Female	223	41.76	206	45.68		6974	41.73		2822	61.08		10106	41.60	
Ethnicity														
Asian	11	2.06	6	1.33		415	2.48		101	2.19		565	2.33	
Black	40	7.49	61	13.53	.0027	2490	14.91	<.0001	357	7.73	.9821	1498	6.17	<.0001
Hispanic	17	3.18	19	4.21		567	3.40		144	3.12		838	3.45	
Native	0	0	0	0		11	0.66		5	1.08		11	0.45	
Other	18	3.37	8	1.77		238	1.43		82	1.77		677	2.79	
Unknown	27	5.06	39	8.65		1303	7.80		482	10.43		3405	14.02	
White	421	77.53	318	70.51	.0016	11677	69.92	<.0001	3449	74.65	.03899	17301	71.21	<.0001

The p-values compare the average values in that group to the average values that we got for the KDS group. In table 19 we see that the BMI values were not statistically different from one another. In table 20 the ethnicity was grouped by white and non-white and in all cases the ethnicity groupings were all significantly different from one another. Similarly, we grouped the ethnicity by black and non-black. We found that all the groups were significantly different except for the KDS group and the OKU group when using this ethnicity grouping. Looking at the summaries we see that there are many more patients in the NCA group and OTH group compared to the KDS, ABP, and OKU groups. There is also a higher presence of males in this data, with more than 50% of the patients being males in each of the categories except for the OKU group. Finally, we notice that there are a lot of white patients represented in this data, which means our data may not fit well in populations with low amounts of white patients.

6.7 Healthy VS Unhealthy Ranges

Another thing that we looked at was how our patient's lab values compared to the normal ranges for these lab values. Because we had people who were admitted to the hospital, we figured that our population would be much different than the normal population. For this we looked at many different variables including pH, bicarbonate, calcium, chloride, hemoglobin, and lipase. Out of the many variables we found a few where most of the patients had unhealthy ranges. These variables include creatine kinase, glucose, lactate, and PO2. For creatine kinase we used the range of 38-174 units per liter for males and 96-140 units per liter for females. We got this information from the blood book webpage [Bloodbook, 2013]. For creatine kinase we found that most patients in each of the five groups had lab values that were lower than the normal or healthy ranges. For glucose we used the range 70-100 mg/dL. We found that most patients in each of the groups had values that were higher than the healthy values. Next, for lactate, we used the range 4.5-19.8 mg/dL. We found that most of our patients in each of the groups had values that were lower than the healthy range. Finally, for PO2, we used 83-100mm Hg as the healthy range. While the numbers we got split between higher values and lower values, an overwhelming number of patients had unhealthy levels. Since these variables seemed to do well categorizing patients between healthy and unhealthy values, we decided to see if any of their regressions would produce better numbers than the regressions we looked at thus far. It turns out that none of these regressions gave us better results than the ones we already used. These regressions produced accuracy values that landed between 0.5 and 0.6, which are much lower than the 0.7 and 0.8 accuracy values we saw when using other variables.

6.8 Distribution Bar Graphs

In addition to logistic regressions, we examined age distributions based on gender, BMI distribution, and EGFR ranges, where the EGFR is calculated using the CKD-EPI equation. The CCI was used in [Chen, et.al., 2018]. The first graphs we will look at are the ones looking at age ranges for each gender. Figure 20 shows the two different bar graphs. In both graphs we can see an increase in the diagnosis of a kidney stone as the age increases. In males however, after about 80 years of age, the rates of kidney stones decrease. In females, the rate of kidney stones stays high as the age continues to increase. Males hit the highest kidney stone rates at the age range of 71-80, while females hit their peak at 71-80, but stay at about the same rate as they get older than 80. For abdominal and back pain, the males follow a similar pattern where it increases, peaks at the range of 61-70, and then starts to decrease. The pattern for females is different from their kidney stone pattern as the rates for abdominal and back pain increase, peak at the range of 51-60 and then decrease. For the NCA, OKU, and OTH groups the male rates increase as they get older until a point where they start to decline. Females, however, have rates that continue to increase as they grow older.

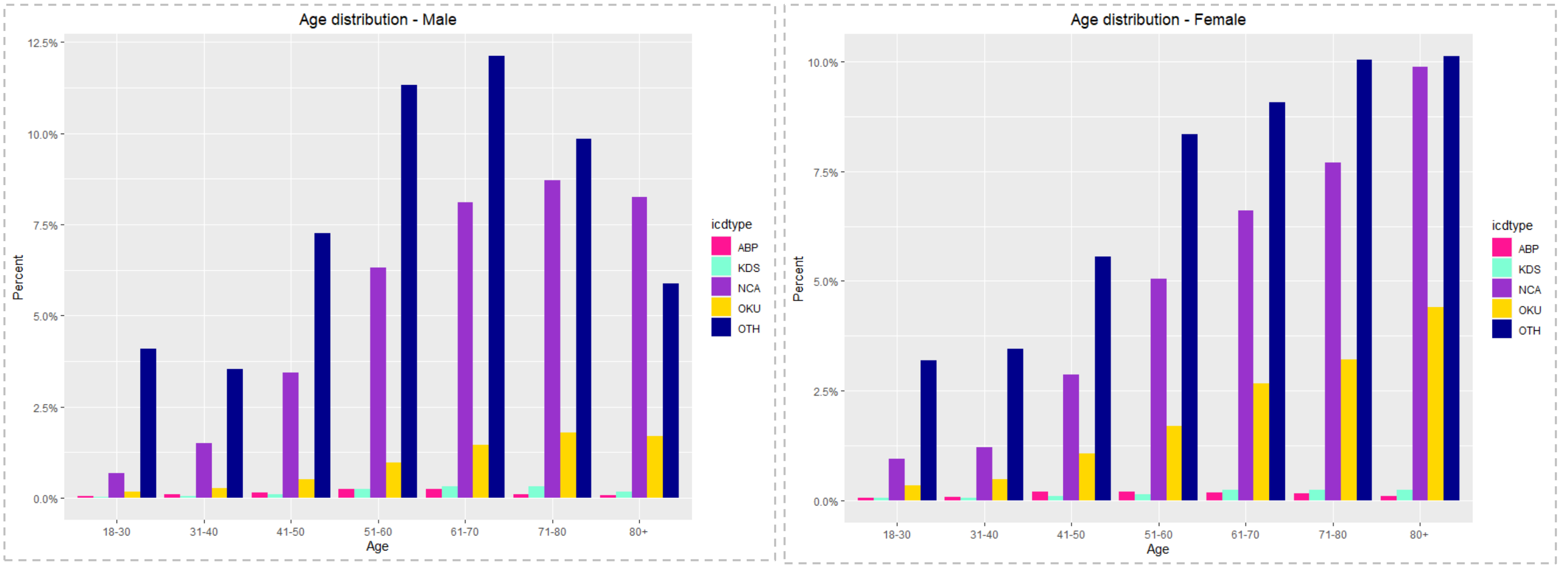


Figure 20 Age distribution of icdtypes based on gender

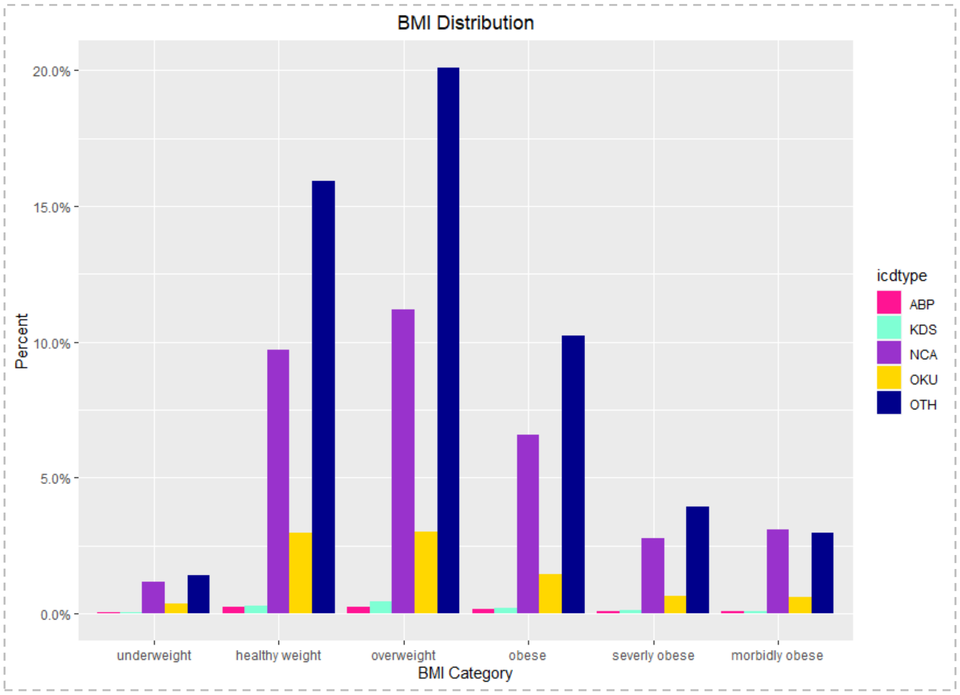


Figure 21 Icdtypes grouped by BMI

In addition to grouping by gender, we grouped the patients in each category by their recorded BMI values. Figure 21 shows the corresponding bar graph. To determine these categories, we looked up the government standards. A BMI less than 18.5 is in the underweight category. A

BMI between 18.5 and 24.9 is considered a healthy weight. A BMI between 25 and 30 is overweight. A BMI between 30 and 35 is obese. A BMI between 35 and 40 is severely obese. A BMI greater than 40 is morbidly obese [BMI]. From this graph we see that the highest prevalence of kidney stones occurs in people who are considered overweight. The category with the next highest prevalence of kidney stones was the healthy weight category. We also notice that the abdominal and back pain group follows the same pattern being most prevalent in overweight patients and then healthy weight patients. We notice that the majority of the NCA and OTH patients have BMI’s that are higher than the healthy range. The OKU group has most of its patients in the healthy weight category or the overweight category.

We then looked at the ethnicity distribution in each of the five groups. Figure 22 shows the corresponding bar graph.

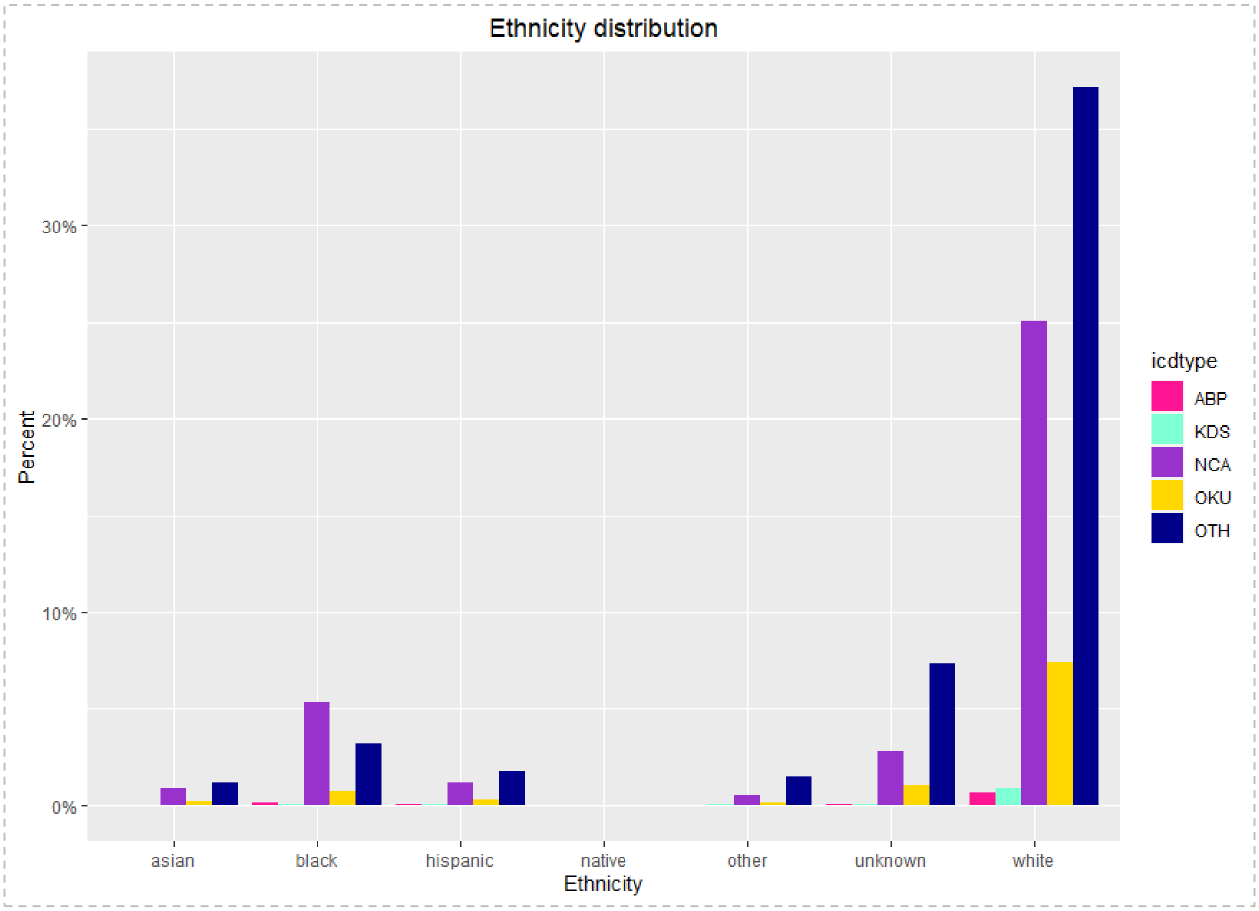


Figure 22 Icdtypes grouped by ethnicity

As we can see in figure 22 there is a huge Caucasian ethnicity majority in all our groups. The next two groups with the most patients are of black ethnicity and unknown ethnicity. Having so many white patients indicates that the results we have found may not apply very well to other ethnicities.

Finally, we grouped the icd types based on EGFR ranges. We used healthline to give us values which indicate how well one's kidney is functioning. There are five different kidney function stages. Stage 1 indicates minimal or no loss of kidney function and is expressed by a

GFR of 90 or above. Stage 2 indicates mild loss of kidney function and is expressed by a GFR of 60 to 89. Stage 3 indicates moderate loss of kidney function and is expressed by a GFR of 30 to 59. Stage 4 indicates severe loss of kidney function and is expressed by a GFR of 15 to 29. Finally, stage 5 indicates kidney failure and is expressed by a GFR of 15 or below [Nall, 2021]. Figure 23 shows this bar graph.

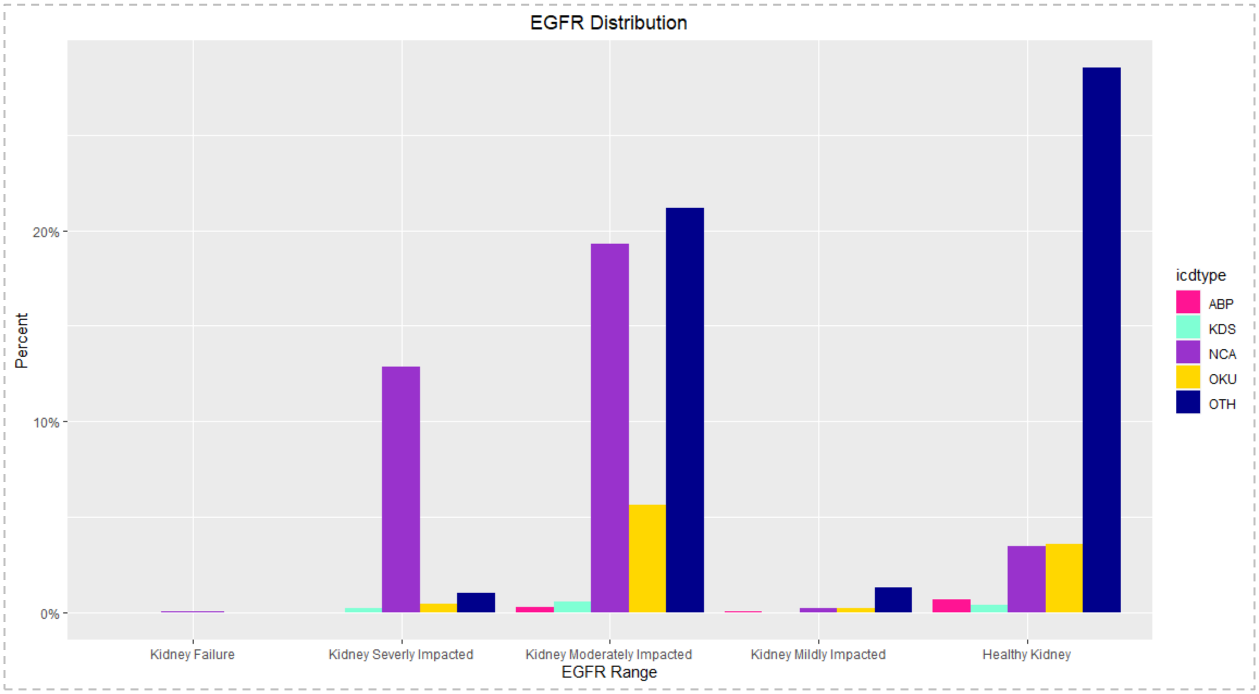


Figure 23 Prevalence of icdtypes grouped by EGFR ranges

From this graph we see that the KDS patients have the highest prevalence in the moderately impacted kidney category. The KDS patients have their second highest prevalence in the healthy kidney category. We see this same pattern with the OKU patients. The OTH and ABP groups have the highest prevalence in the healthy kidney category followed by the moderately impacted category. The NCA patients are different from all the other groups as nearly all their patients have a moderately impacted kidney or a severely impacted kidney. Overall, most of the patients have either a healthy kidney or a moderately impacted kidney.

7. Discussion

Our results gave us some interesting perspectives when it comes to kidney stones and diseases related to the kidneys or the urinary tract. We were encouraged that our results, although not the best, were very similar to the results of the diagnostic acute care algorithm for kidney stone disease [Chen, et.al., 2018]. We found through our analysis that when comparing the KDS group VS the ABP group a logistic regression using the five variables age, chloride, creatinine, resprate mean, and CO2 levels from the data set that contained the patients first lab results gave us the best accuracy in distinguishing these two groups. The accuracy for this regression was 0.699. This also had a specificity of 0.661 and a sensitivity of 0.726. For the KDS group VS the NCA group we found that a logistic regression using the Elixhauser score and the BUN values with the data that contained the first lab results from the first admittance produced the best outcome. This outcome had an accuracy of 0.883. The specificity for this had a value of 0.898 and a sensitivity of 0.508. For the KDS group VS the OKU group we found that a logistic regression using the EGFR calculated from the CKD equation with the average lab values gave us the best outcome. This outcome had an accuracy of 0.852. It also had a specificity of 0.922 and a sensitivity of 0.247. Finally, for the KDS group VS the OTH group, we found that a logistic regression using age, BUN, creatinine, EGFR, and CO2 gave us the best outcome. This outcome had an accuracy of 0.894, a specificity of 0.903, and a sensitivity of 0.323. In the case of KDS VS ABP we were able to maximize the sensitivity, meaning that this model is best used when trying to identify patients with kidney stones. In all the other cases we were able to maximize the specificity, which means that these models do best when trying to identify patients without kidney stones. Some possible problems with our results come from the data we used. First, we notice that there is a relatively small amount of data from patients with kidney stones as most of the patients in this data were not diagnosed with kidney stones. Also, we had a high presence of males in our overall population. This indicates that our results could have a bias towards a high male population and may not do well in populations high in females. We also see that most of the patients had either a black or Caucasian ethnicity. This indicates that our results may not work very well in populations that are not dominantly black and Caucasian. Sample R codes that we used for this research can be found in appendix 5.

Appendix

Appendix 1: Learning R and Key Tools for Kidney Stone Research

1.1 Learning Coding Through R

Initially, it was necessary to download R before it could be used. To do this we went to the R project website [R project]. After installing and downloading this program we installed Rstudio, which makes R more user friendly. After installation of R and R studio we needed to learn how to use it because R is the program that we used to create models and evaluate their accuracy. Doing a little bit of research, we were able to find a website that provided lots of information on what the system R is and how to do basic codes like means, averages, and graphs [R Tutorial, 2020]. After learning some of the basic codes in R, Statquest videos provided extra information about what the numbers received in R meant. Topics like logistic regression, correlation, and random forests were discussed to provide a better understanding of what the R output truly meant. Reading and watching these things were helpful, but to truly understand these things they must be performed. Using a subset of the data, we put into action what we had learned from the readings and videos. This data included variables, like the patient's age, gender, ethnicity, and their icdtype which was either kidney stones (KS) or abdominal and back pain (ABP). This subset of data was very useful for learning the ropes of R as well as determining some factors that may be important. This subset of data gave us a strong foundation to build on when we started to work with the entire data set. Additional learning work was done using heart disease data. This information and some of the results can be found in appendix 2.

1.2 Charlson Comorbidity Index (CCI)

The Charlson Comorbidity Index is a prediction metric that predicts the ten year mortality rate of someone presenting one or more conditions in the model [Ciorniciuc, 2015]. The index includes the patient's age and 16 other possible conditions. Some of these other conditions include myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, and dementia [Ciorniciuc, 2015]. A patient with one or more of these diseases is given a score of 1, 2, 3, or 6. The score is based on which category the disease is in. A disease scored with a value of 1 has a lower likelihood of leading to death, while a disease scored with a value of 6 has a high likelihood of death. Some of the diseases scored as 1 include connective tissue disease, ulcer, and chronic liver disease. Some diseases that are scored as a 2 are solid tumor, leukemia, and lymphoma. The only three point disease is moderate to severe liver disease. Finally, the three diseases that are scored a 6 are malignant tumor, metastasis, and AIDS. Adding up all the values gives the CCI score. To calculate the survival rate of the patient we use the formula $0.938^{(e^{(C*0.9)})}$ where C is the CCI score [Ciorniciuc, 2015].

1.3 Stone Score and Stone Plus

The stone score is used as a predictor of a patient being stone free and complication free after percutaneous nephrolithotomy (PCNL), which is a kidney stone removal surgery [Jiang, et. al., 2019]. Stone stands for stone size, tract length, obstruction, number of involved calices, and essence [Jiang, et. al., 2019]. Obstruction refers to hydronephrosis, which is when the kidney swells due to improper drainage of urine [Stephens, 2020]. Essence defines the stone density in Hounsflied units [Nedea, 2017]. To calculate the stone score, certain conditions are given a certain amount of points. For stone size, a size of 0-399mm^2 is given one point, a stone of size 400-799mm^2 is given two points, a stone of size 800-1599mm^2 is given three points, and a stone >1600 mm^2 is given four points [Nedea, 2017]. A tract length less than or equal to 100mm is given one point and anything above 100mm is given two points. For obstruction, no/mild hydronephrosis is given one point and moderate/severe hydronephrosis is given two points. If the number of involved calices is 1 or 2 then one point is given. If there are 3 calices, then two points are given. A full staghorn stone is given three points. Finally, an essence less than or equal to a value of 950 is given one point and an essence greater than 950 is given two points [Nedea, 2017]. Table 1 shows the probability of someone being stone free based off their stone score.

Table 1 Stone free percentages based on stone score [Nedea, 2017]

STONE result	**Probability of stone free outcome after first procedure**
5	94%
6	92%
7	88%
8	83%
9	64%
10	42%
≥11	27%

Another way to calculate a STONE score is by using sex, duration of pain, race, nausea and vomiting, and hematuria. This method is used more to determine whether someone has a kidney stone. Similarly, like the other method, points are assigned for certain categories. For sex, males receive two points and females receive no points. A duration of pain greater than 24 hours receives no points, 6-24 hours receives one point, and less than 6 hours receives three points. An ethnicity of black receives no points and non-black receives three points. For nausea and vomiting, not present receives no points, only one present receives one point, and when both are present two points are given. Finally, if hematuria is present three points are given and if it is absent no points are given [Safaie, et. al., 2019]. To interpret the numbers, a point total of 5 or less means a low risk of having kidney stones, 6-9 means a moderate risk of having a kidney stone, and a score of 10 or greater means a high risk of having a kidney stone [Moore, 2020]. Stone plus is an addition to this second method of calculating a stone score. Plus refers to point of care limited ultrasonography, which is an ultrasound done at the patient's bedside [Daniels, et. al., 2016]. Stone plus combines the stone score mentioned above and the bedside ultrasound to predict someone’s chances of having a kidney stone more accurately. This ultrasound can be used to determine if a patient has hydronephrosis [Daniels, et. al., 2016]. The presence or not of hydronephrosis is then factored into the stone score and could possibly change one’s likelihood of having a kidney stone [Daniels, et. al., 2016]. The whole idea behind the stone plus is to improve accuracy of diagnosing someone with or without a kidney stone without having to cause unnecessary imaging radiation exposure.

1.4 JESS and Supersaturation

JESS stands for Joint Expert Speciation System. JESS is a research tool that is used to model chemical speciation in aqueous environments [May, et.al., 2019]. Chemical speciation is the process of identifying and quantifying the abundance of a chemical or ion in a certain aqueous solution [May, et.al., 2019]. The JESS interface contains over 12,000 thermodynamic reactions, which includes chemical reactions and their parameters such as equilibrium constants and reaction enthalpies [May, et.al., 2019]. Other factors that are considered are amounts of components, ionic strength, temperature, and solution ph. To model a certain reaction the information about the chemical reaction is taken from the JESS database and “is then transformed into a thermodynamically consistent set of equations and solved” [May, et.al., 2019]. JESS models are much more useful to predict chemical behavior or abundance then they are for providing an analytical point of view [May, et.al., 2019]. This can be applied to all chemical species present in urine, but calcium oxalate and calcium phosphate are of increased importance because they lead to kidney stones [Rodgers, et. al., 2014]. Therefore, JESS can be used to predict the supersaturation of key compounds related to kidney stone formation, like calcium oxalate, calcium phosphate, and uric acid, which are found in urine. Supersaturation of a liquid refers to when a solution has more dissolved solute than it should at that volume or temperature [Supersaturated, 2020]. Supersaturation is important because supersaturation is key to crystal nucleation, which is the first step in the formation and growth of kidney stones. Being able to predict the supersaturation of compounds in urine can help predict the likelihood of a kidney stone forming in a patient. In addition, it can help determine which type of stone will form, allowing doctors to prescribe certain remedies that can help prevent the patient from getting a kidney stone of that type. A couple limitations of JESS are that it does not consider kinetic phenomena and its accuracy is limited to that of its database of thermodynamic constants [May, et.al., 2019]. Fortunately, the JESS database has over 12,000 thermodynamic constants. While the thermodynamic constants limit the accuracy, they also make the results reliable [May, et.al., 2019]. JESS is commonly used on astronauts when they return from space because the pressure and diet change, while in space, changes the urine chemistry of the human body [Myers, et. al., 2018].

Appendix 2: Logistic Regression from Data Subset

With the kidney stone data subset, we wanted to get some practice doing regressions and random forests, so we looked at the variables that seemed to be correlated with icdtype. Initially, a logistic regression was done just using age. In other words, age was used as the only predictor for determining if someone is likely to have kidney stones. Figure 1 shows the logistic regression curve and the corresponding probability-curve using equation 1, given below.

$$\log(\text{odds}) = (-0.1903) + 0.0055(\text{age}) \quad \textbf{(1)}$$

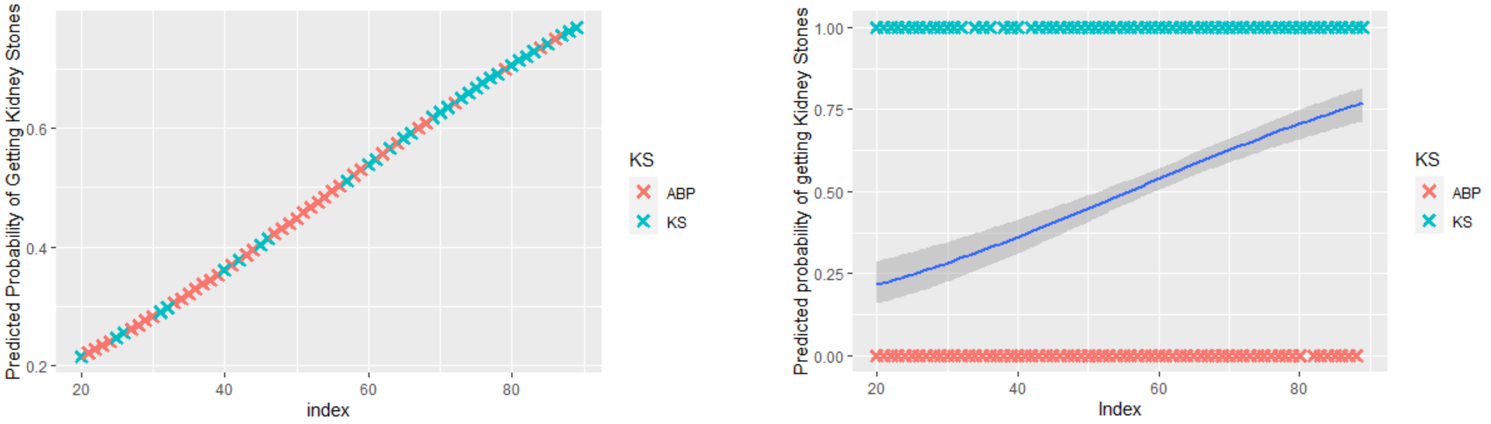


Figure 1 Logistic regression(left) and corresponding probability-curve (right) for equation 1

The index is simply just the age of the patients. We note that any patient who has an age greater than 89 corresponds to the value 300 and have been omitted from both graphs of figure 1. Looking at the left graph of the figure we can see that all the patients in this set of data have at least a 20% chance of developing a kidney stone. There is somewhat of a divide between patients with kidney stones and those without. We can see that at about .5 on the y-axis most of the points above that line have kidney stones while the majority below that do not. However, we do see a good amount of overlap with patients who don't have kidney stones above that line and vice versa. The very low coefficient for age tells us that age does not have much of an impact on determining kidney stones as its coefficient is very close to 0. As the p-value for age is significant at a value of .0014, age does however help predict kidney stones, but helps at a very minimal level.

We wanted to look at the validity of this logistic model by looking at the binned residual plot. Figure 2 shows this residual plot. The grey lines represent the 95% confidence interval. The residual plot looks very good. All the points except for one lie within the 95% confidence interval. This tells us that this model is very reasonable to use. There do seem to be more points that have a negative value, which would indicate that the model is more likely to over predict. Overall, the residual plot does not raise any concerns about the validity of the logistic model.

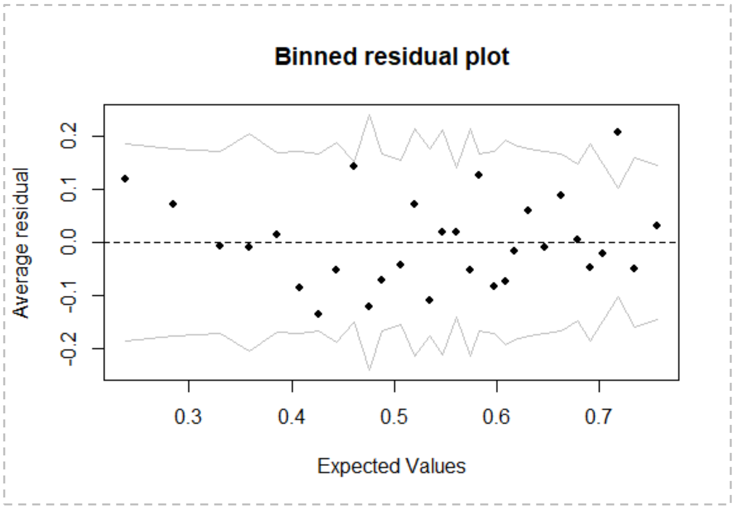


Figure 2 Binned residual plot using equation 1

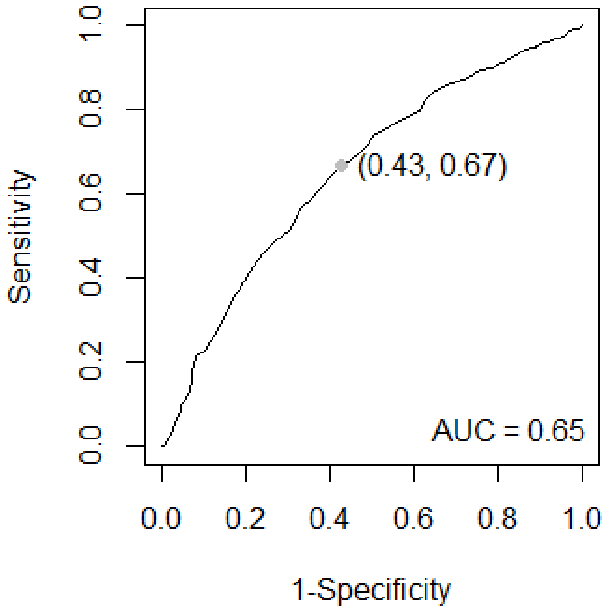


Figure 3 ROC curve for logistic regression using just age

Figure 3 shows the ROC curve for the logistic regression that just uses age. In the data that we have, there are 445 patients in the ABP group and 534 patients In the KS group. The area under this curve is 0.6536. Ideally, we want this number to be 1. Instead we get a number that is close to 0.5. This indicates that this regression model is a bit better than randomly assigning a patient to a category.

We ran a random forest, which had 500 trees, using just age to predict kidney stones and got the results shown in table 2.

Table 2 Confusion matrix for random forest using equation 1

	KS	ABP
KS	372	162
ABP	223	222

For this random forest we got an OOB (out of bag) estimate of error of 39.33%. This tells us that 61.67% of the out of bag samples were correctly classified by the random forest. The value of the class error for the second row was .501, which tells us about 50% of patients predicted to have abdominal or back pain (ABP) were incorrectly identified. The first row had a class error of .303, which tells us about 30% of people with kidney stones (KS) were incorrectly identified. These large error rates indicate that age alone isn't very good at predicting KS and ABP.

We then did a similar thing looking at both age and gender as predictors for kidney stones. To do this we needed to make all males have a value of 1 and all females have a value of 0. This logistic regression gave us equation 2.

log(odds)= -0.2697+ 0.0050*age + 0.2083 (if male) **(2)**

Again, age had a significant p-value of 0.002, but gender did not have a significant p-value, which had a value of 0.11. This tells us that age has a very small effect on predicting kidney stones, while gender possibly has no effect on predicting kidney stones. For both variables, the coefficients are small and so neither one of them has a large impact on the prediction of having kidney stones. Figure 4 shows the graph for the logistic regression and the corresponding probability-curve for equation 2. The index is represented by the rank, which is ordered by the probability of kidney stones from lowest to highest. So, 1 corresponds to the data point with the lowest probability of getting a kidney stone and 2 is the data point that has the second lowest kidney stone probability, and so on.

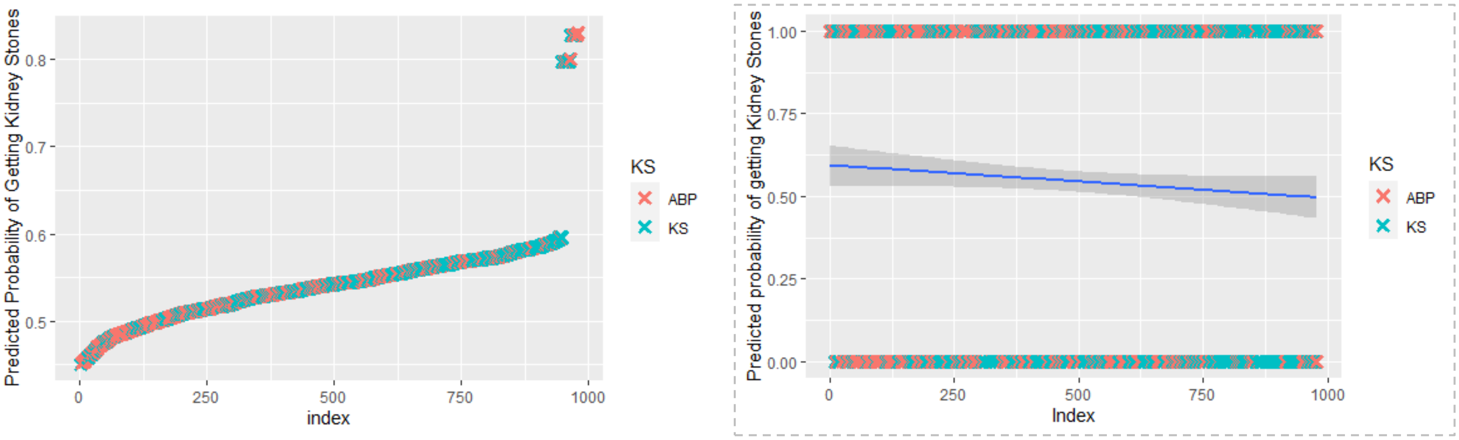


Figure 4 Logistic regression (left) and corresponding probability-curve (right) for equation 2

Similarly, in this graph, we see that all patients have about a 0.5 probability of having kidney stones, with only about 125 of the patients having less than a probability of 0.5. Also, in this graph more than the other, there is not a clear separation between kidney stone patients and abdominal and back pain patients, further showing how this is not the best model for predicting kidney stones.

We decided to also look at the binned residual plot for this logistic regression model. The binned residual plot is shown in figure 5. The grey lines represent the 95% confidence interval. This plot again looks very good. There is only one point that lies outside of the 95% confidence interval. This plot indicates that the logistic model using equation 2 is reasonable. In addition, we see that the points seem to be closer to the zero line as the expected values increase. This indicates that this model does better predicting at higher expected values than the lower values.

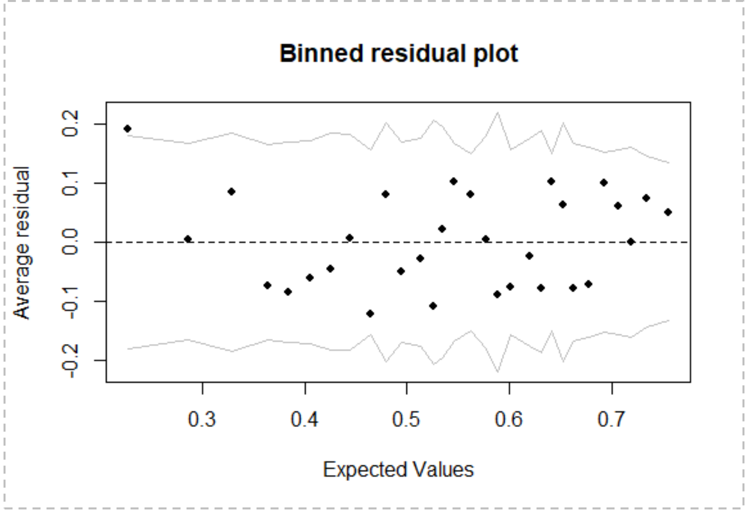


Figure 5 Binned residual plot using equation 2

Figure 6 shows the ROC for the logistic regression that uses both age and gender as predictors. The AUC for this graph was 0.613. This is not great as it is close to 0.5. This means that this regression is just a bit better at categorizing patients than just random selection. Comparing the two ROC graphs and AUC values we see that the model that just uses age is slightly better when it comes to predicting patients with kidney stones than the model that uses both age and gender.

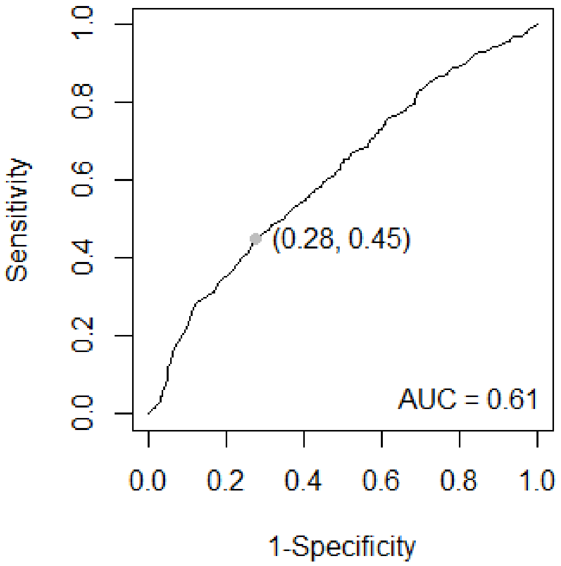


Figure 6 ROC graph for logistic regression using equation 2

We again did a random forest using these two variables. The results are shown in Table 3.

Table 3 Confusion matrix for random forest using equation 2

	KS	ABP
KS	405	129
ABP	247	198

For this random forest we got an OOB (out of bag) estimate of error of 38.41%. This tells us that 62.59% of the out of bag samples were correctly classified by the random forest. The value of the class error for the second row was 0.555, which tells us about 56% of patients predicted to have ABP were incorrectly identified. The first row had a class error of 0.242, which tells us about 24% of people with KS were incorrectly identified. These again aren't great numbers indicating that these two variables may not be the best when it comes to predicting kidney stones.

Appendix 3: Other Experiments: Classification and Prediction Practice

3.1 Data Set Description

The heart disease data includes information from 299 patients with heart failure that was collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad in 2015 [Chicco and Jurman, 2020]. Machine learning was applied to predict a patient's survival rate after heart failure, using many variables like age, sex, ejection fraction, serum creatinine, and anemia. The patients included 105 women and 194 men that were between the ages of 40 and 95 [Chicco and Jurman, 2020]. In total there were 13 variables in this data set that were binary or numerical variables. Some definitions are needed for serum creatinine, ejection fraction, and death event. Serum creatinine is a byproduct of a biological function that breaks down muscle tissue [Chicco and Jurman, 2020]. Ejection fraction is an indication of how well your left ventricle pumps blood throughout your body. It looks at the total percentage of the blood that leaves the left ventricle [Mankad, 2019]. Finally, the death event refers to the patient's state, whether dead or alive, before the end of the follow-up period which was an average of 130 days [Chicco and Jurman, 2020].

3.2 Methods

Davide Chicco and Giuseppe Jurman made all their codes available for public use. To get access to these codes we downloaded the data set and the software code that was provided by them under the availability of data and materials section of the paper [Chicco and Jurman, 2020]. Then we were able to load this information into R and use the code that was provided. Changing minor things like the location of the materials in the folder or plotting the graphs, we could simply run the code and produce the exact same graphs that they did.

3.3 Results

The most important variables for determining heart failure are serum creatinine and ejection fraction when considering one's survival rate from heart failure. Figure 7 shows the aggregated results of the feature rankings. This is a summary of many different feature rankings that were done with the data. Some of the feature rankings that were done were chi-square feature rankings, and the Gini index.

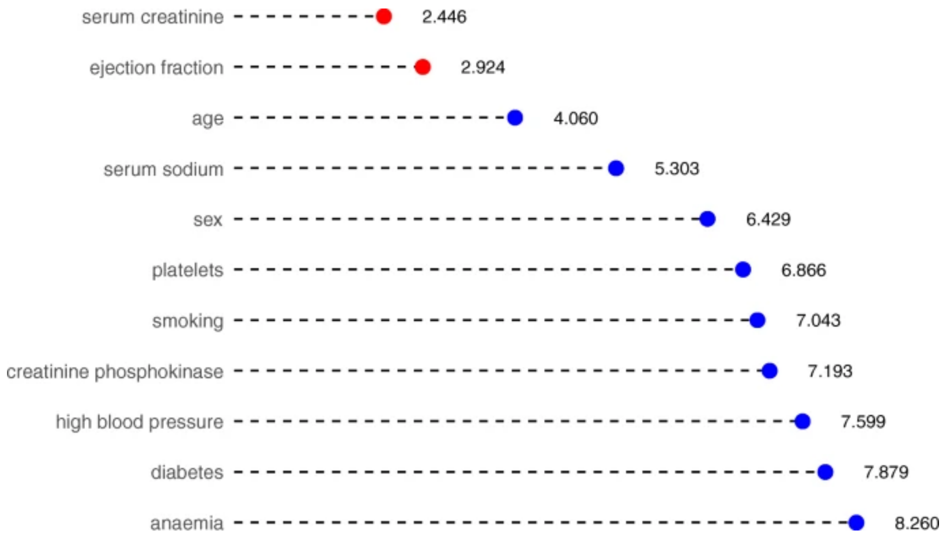


Figure 7 Summary of feature rankings [Chicco and Jurman, 2020]

In this case a lower number means a higher importance for the feature. As serum creatinine and ejection fraction has the lowest numbers; they are the most important features followed by age and serum sodium. Figure 8 shows the final scatterplot of serum creatinine and ejection fraction.

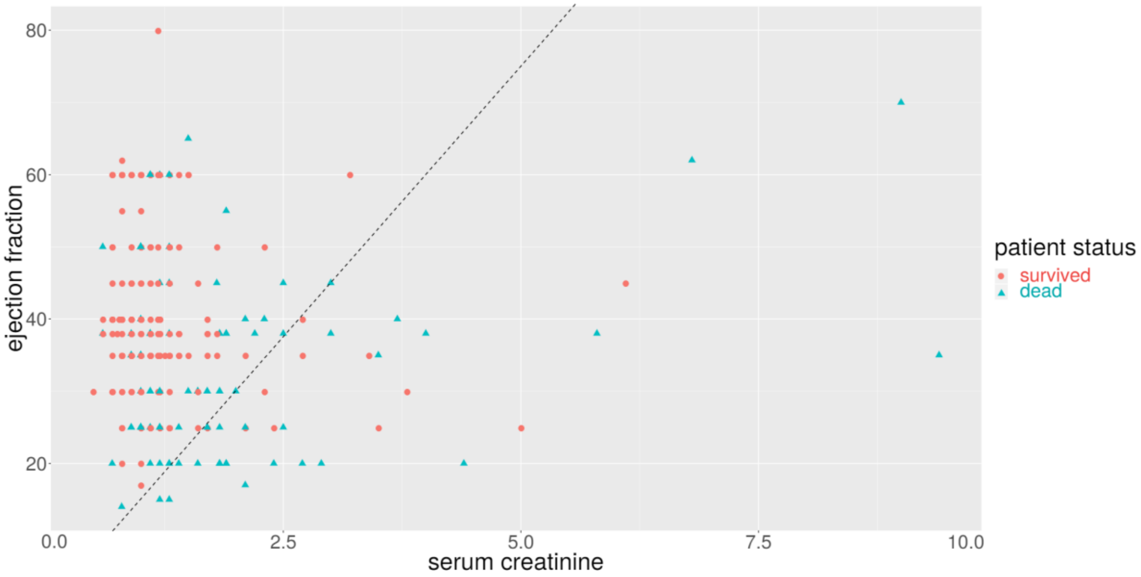


Figure 8 Scatterplot for ejection fraction and serum creatinine [Chicco and Jurman, 2020]

Based on this plot and since we are talking about the death or survival of the patient, we feel as though this does not do a very good job discriminating between alive and dead patients. This is because there is a lot of overlap especially with the dead patients. A lot of patients that were predicted to be alive, had died. Our viewpoint does not agree with the authors of the paper, who say this graph “shows a clear distinction between patients who are dead and patients who are alive” [Chicco and Jurman, 2020]. While this scatterplot uses the best predictors, we feel as though they don’t do an exceptional job in predicting.

Another graph used by Chicco and Jurman was the survival percentage after a certain number of months after heart failure. Figure 9 shows the bar graph of the survival rates. We notice that there is a massive jump in survival chances if a patient survived past 1 month after heart failure. The survival rate goes from a less than 25% chance to almost a 75% chance of survival. We also see that five months seems to be an outlier in this data as the survival rate after 2 or more months is about 75% or greater except for five months after heart failure, which only has about a 45% survival rate.

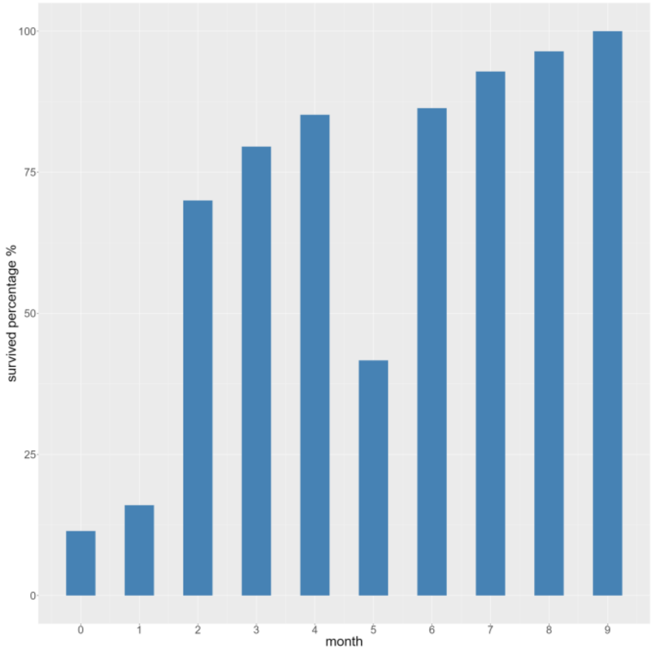


Figure 9 Survival percentage after x-months [Chicco and Jurman, 2020]

3.3.1 Logistic Regression and Random Forests

The first thing that we did was look at the logistic regression that was formed from just using ejection fraction. The equation that was formed is given in equation 3.

$$\log(\text{odds})=1.31+(-0.056)\ (\text{ejection fraction}) \qquad \textbf{(3)}$$

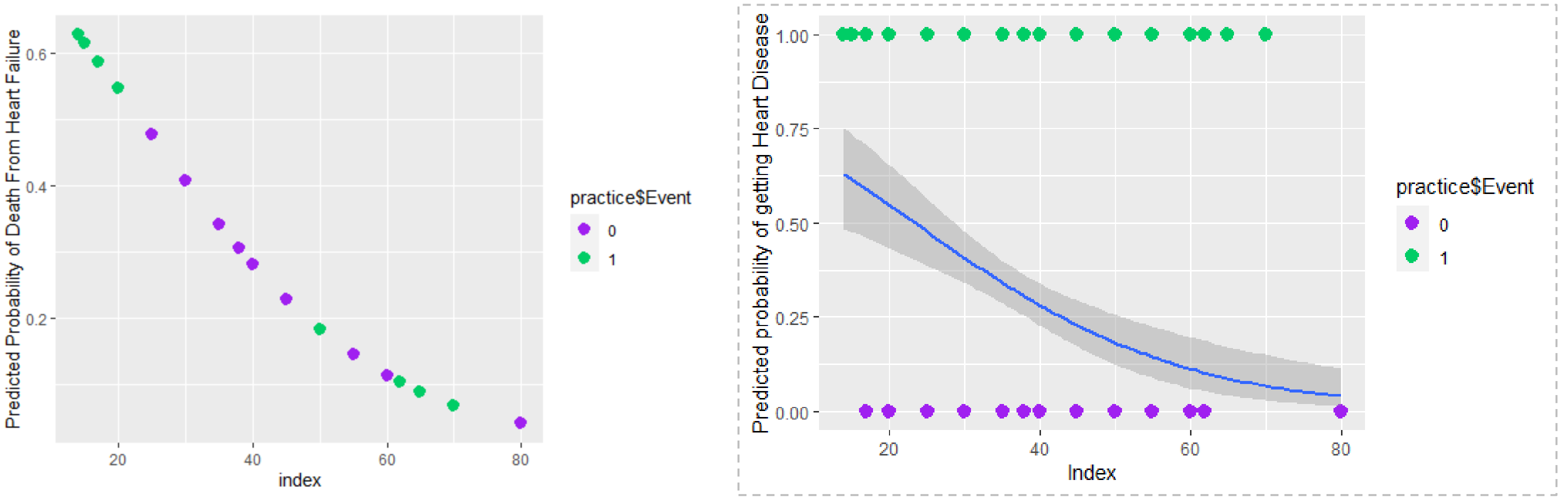


Figure 10 Logistic regression (left) and corresponding probability-curve (right) for equation 3

For this logistic regression, ejection fraction had a coefficient of -0.056 and a very significant p-value of $7.88 * 10^{-6}$. This indicates that ejection fraction influences one's survival rate after heart failure even though it is a very small effect. The logistic regression and corresponding probability-curve that was produced by equation 3 is shown in figure 10. The index represents the value for ejection fraction. There is a clear trend that shows the chances of dying from heart failure decreases as the ejection fraction increases. However, this probability-curve isn’t super helpful since we have data points for each category scattered along the x-axis. We don’t have a pattern where all who survived have low index values and all who died had high index values. This makes it very difficult to classify who will die from just ejection fraction alone. After creating a ROC curve, we got an AUC value of 0.6761, which isn’t bad as we want the number to be as close to 1 as possible. Using 100 executions for training data and testing data, where a random 20% of the data is used as testing data for each execution, we were able to calculate the mean MCC, true positive rate, true negative rate, accuracy, and the AUC. For the average MCC we got a value of 0.324, which is closer to 0 than we would like, as 0 represents the graph being just as good as random assignment. For the true positive rate, we got a value of 0.835, which indicates that on average 83.5% of the true cases were correctly identified. Our true negative rate was 0.724, which is a good value, but not great in a medical situation, like heart disease. The accuracy value was 0.726, which means this model has a mean accuracy of about 72%. Finally, we got a mean AUC value of 0.781.

As we did not get the best results from the logistic regression model, we wanted to see if this model was a reasonable model to use. To do this we looked at the residuals versus predicted and made a binned model, which averages the residual values. Figure 11 shows the binned model for the residuals vs predicted using equation 3.

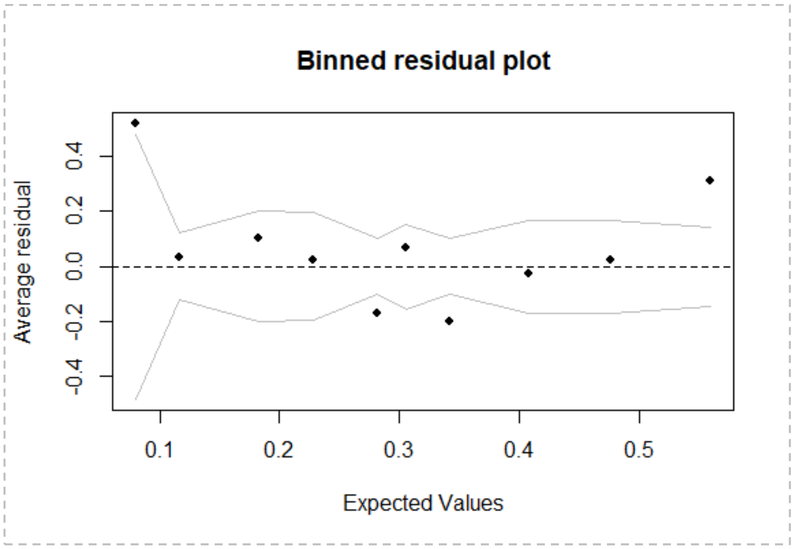


Figure 11 Binned residual plot using equation 3

The grey lines in figure 11 represent the 95% confidence interval for the average residuals. Looking at this figure we see that the model is not great as four of the points lie outside of the 95% confidence interval. Since there are not that many points overall, having four points outside of the interval indicates that there are more outliers in this data than we would normally expect. It looks like this model does a good job predicting from 0.1-0.25 and 0.4-0.5 but does not do a good job at the other intervals. Overall, this plot would indicate that this logistic model is reasonable to use for this data.

We also did a random forest that uses equation 3. The results of this are shown in table 4.

Table 4 Confusion matrix for logistic regression using equation 3

	1 (patient died)	0 (patient survived)
1 (patient died)	20	76
0 (patient survived)	16	187

For this we got an OOB (out of bag) error rate of 30.77%. This means 69.23% of the out of bag samples were correctly identified by the random forest.

We wanted to compare ejection fraction to creatinine and did this by doing a logistic regression just using creatinine. This logistic regression was done by using equation 4.

$$\log(\text{odds}) = -1.89+0.8242(\text{creatinine}) \qquad \textbf{(4)}$$

The coefficient for creatinine from the logistic regression was 0.8242 with a very significant p-value. This indicates that creatinine has a significant effect when predicting death from heart failure. Figure 12 shows the logistic regression and corresponding probability-curve using equation 4. The index on the graphs shown in figure 12 represents the creatinine levels.

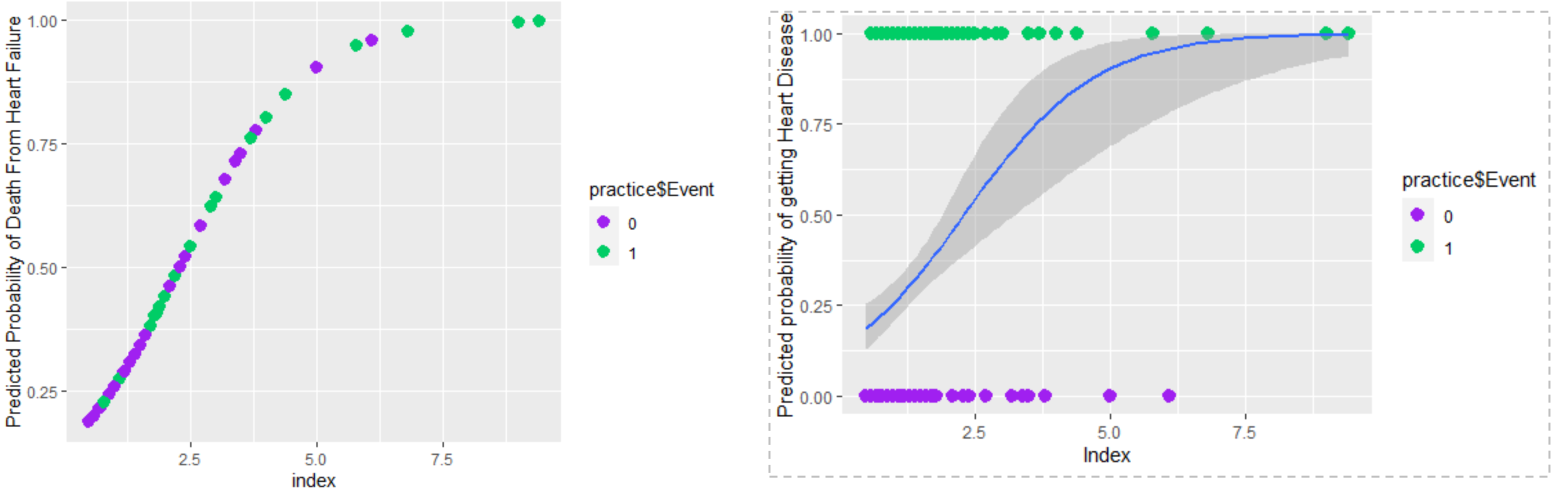


Figure 12 Logistic regression (left) and corresponding probability-curve (right) for equation 4

This graph shows a direct relationship. As the creatinine level increases the probability of dying from heart failure also increases. There does not seem to be a clear separation, which indicates that the level of creatinine alone is not a very good separator when predicting death from heart failure. We created a ROC graph for this regression, and we got an AUC of 0.7281, which is a pretty good value. Similarly, we again did 100 repetitions of testing sets and training sets just using equation 4. For the average MCC value, we got a value of 0.215. This is not a good value to have for MCC and is closer to 0 than the one with just ejection fraction was. This indicated that this regression is even closer to just random assignment. The true positive rate had a value of 0.652, which again is lower than just ejection fraction. The true negative value was 0.709. The accuracy had a mean value of 0.702. Finally, the mean AUC value was 0.681. All the mean values that we got using just creatinine were lower than the ones we got with ejection fraction. This would indicate that ejection fraction is a better predictor than creatinine when looked at individually. This differs from the ranking features in figure 8, which says creatinine is the more important variable, over ejection fraction. We likely see this because both the feature rankings and the training data used 100 executions, which chooses randomly. So, every time this is done, different results would be obtained. That is, if we were to redo the training data for both variables, we could get an outcome that agrees with the feature rankings shown in figure 8.

Similarly, as before, we decided to look at the residuals for the model that uses equation 4. The binned residual plot for equation 4 is shown in figure 13. Again, the grey lines represent the 95% confidence interval for the residual values.

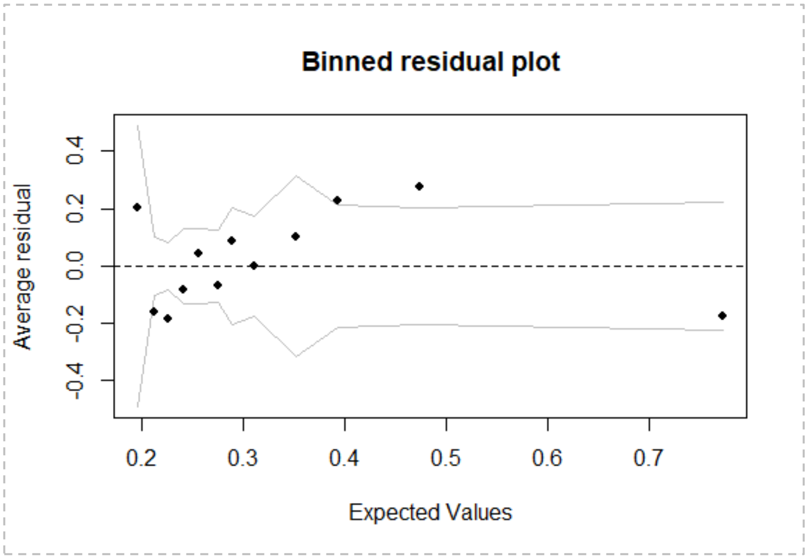


Figure 13 Binned residual plot using equation 4

This residual plot does not look as good as the one for ejection fraction. This plot seems to indicate that the logistic model that uses equation 4 is good at predicting when the level of creatinine is between the expected values of .25 and .35. Any other time it seems that this model does not do a very good job at predicting. This indicated that it probably is not a good idea to use just creatinine when predicting someone's chances of death from heart failure.

We again created a random forest. The results are shown in table 5.

Table 5 Confusion matrix from random forest using equation 4

	1 (patient died)	0 (patient survived)
1 (patient died)	24	72
0 (patient survived)	19	184

This gave us an OOB estimate of error rate of 30.43%, which means 69.57% of the OOB samples were correctly identified by the random forest. These two values are very close to what we got using just ejection fraction, which means the two random forests perform similarly.

Because age was the third highest variable, in terms of importance, we decided to look at its s-curve, do training and testing data, and make random forests. This allowed us to compare it to creatinine and ejection fraction which were the top two variables. Figure 14 shows the s-curve that corresponds to the logistic regression given by equation 5.

$$\log(\text{odds}) = -3.654+0.047(\text{age}) \qquad \textbf{(5)}$$

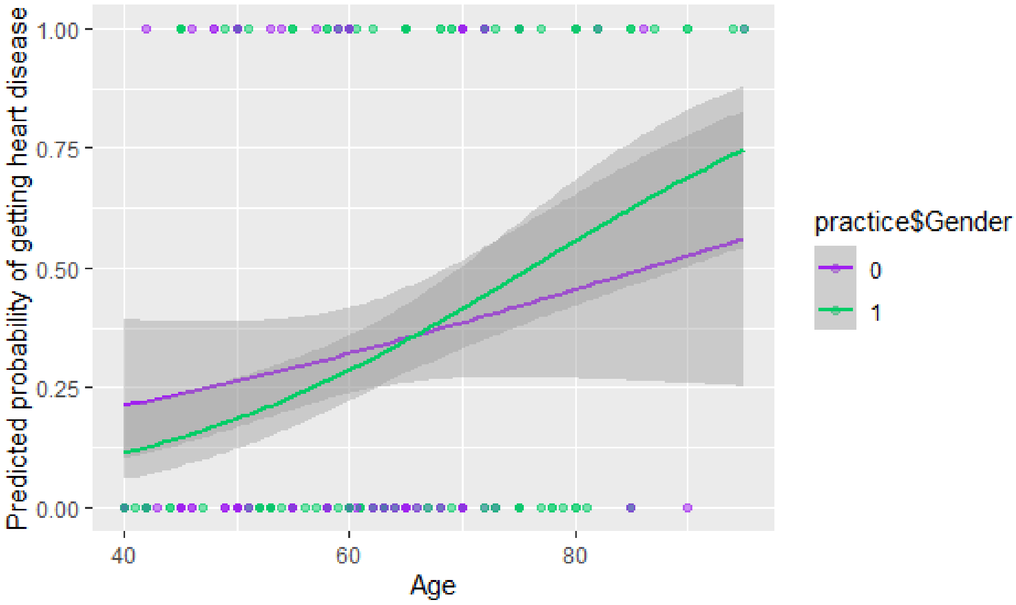


Figure 14 S-curve for equation 5

In this figure the two different s-curves represent the two genders. The purple line and points represent males and the green line and points represent females. As we see in figure 14 both genders are at increased risks of dying from heart disease as they get older. However, based on this data it seems that men have a quicker growth rate in death from heart disease as they get older, while the women's growth rate increases at a slower, steadier rate. We also notice that men start at a lower risk of dying from heart disease than women do, but this changes around the age of 65. Running training data, we found that the average MCC value for this regression is 0.218. The true positive rate was 0.655, while the average true negative rate was

0.718. The average accuracy was 0.709 and the average AUC was 0.686. This regression came out with very similar numbers as the regression that just uses creatinine.

Just like earlier, we again looked at the binned residual plot for the model that uses equation 5. The grey lines again represent the 95% confidence interval for the residual values. Figure 15 shows the binned residual plot. This residual plot looks pretty good. There are, however, some places in which this model seems to do poorly. These places occur at the expected values lower than 0.2 and between 0.3 and 0.4. In these cases, the points lie significantly below the 0 line and outside of the 95% confidence interval. Since residuals are observed minus fitted values, residuals that have a negative value indicates that the model is over predicting at those points [Webb, 2017]. Thus, the model that just uses age seems to overpredict at those expected values.

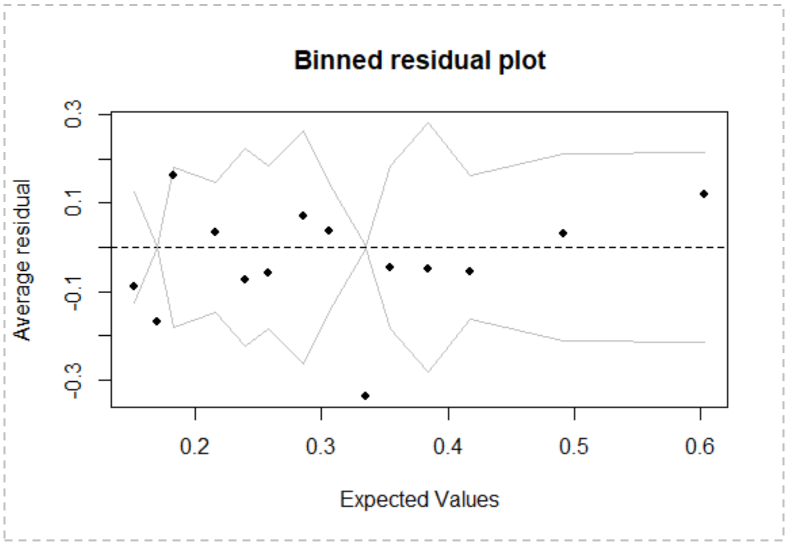


Figure 15 Binned residual plot using equation 5.

We also tried random forests using just age and table 6 shows the results.

Table 6 Confusion matrix from random forest using equation 5

	1 (patient died)	0 (patient survived)
1 (patient died)	27	69
0 (patient survived)	20	183

This gave us an OOB estimate of error rate of 29.77%, which means 70.23% of the OOB samples were correctly identified by the random forest.

Next, we decided to see how the two best factors would do together. Doing a logistic regression, we found that again both creatinine and ejection fraction were important factors in determining death from heart failure. Figure 16 shows the logistic regression curve and corresponding probability-curve for equation 6.

$$\log(\text{odds})=0.378+0.749(\text{creatinine})+(-0.598)\ (\text{ejection fraction}) \qquad \textbf{(6)}$$

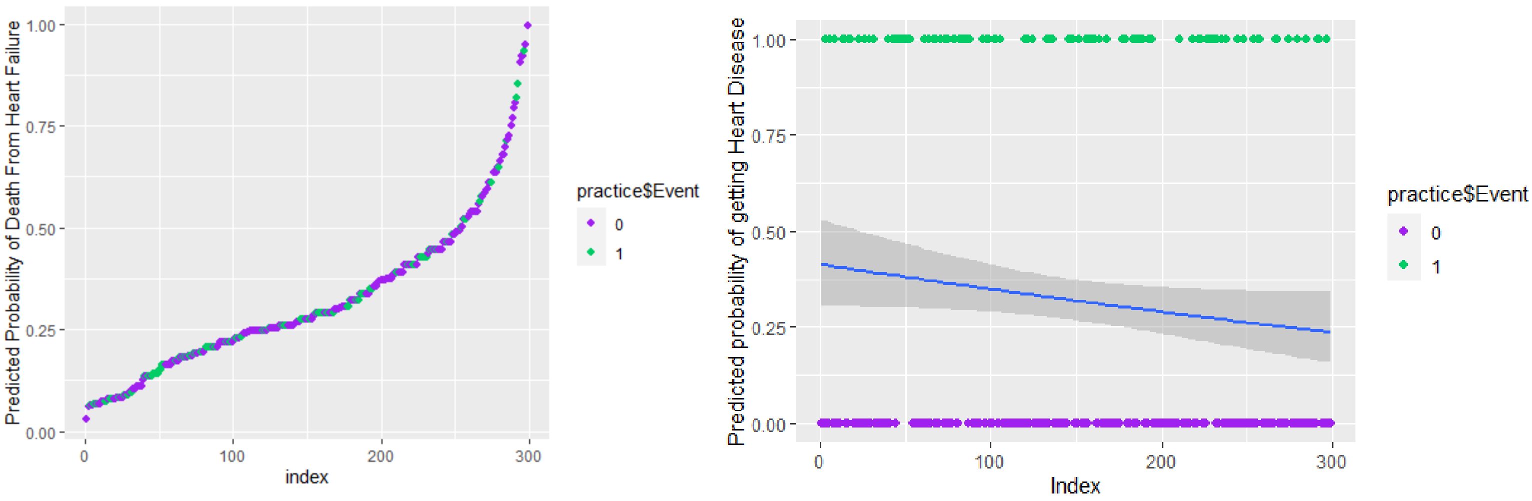


Figure 16 Logistic regression (left) and corresponding probability-curve (right) for equation 6

The index for this graph is based on the rankings of all 299 samples. We do not see a very good result from either one of these graphs. The logistic regression curve has dead patients and survivors mixed throughout and does not separate the patients very well. We see a similar thing in the probability-curve graph. Doing a ROC and finding the AUC for this logistic regression we found the AUC to have a value of 0.7614. We again did 100 executions using randomized training and testing sets. For the model using equation 6 we got a MCC of 0.357, which is the best that we have seen so far. We got a true positive rate of 0.707 and a true negative rate of 0.757. The mean accuracy was 0.749 and the AUC value was 0.732. While the AUC and true positive rate are slightly lower than the ones using just ejection fraction, all the other values were higher. From this information it is safe to conclude this is the best of the 4 models we have looked at so far.

Like before, we again looked at the binned residual plot using equation 6. Figure 17 shows the binned residual plot and the grey lines again represent the 95% confidence interval.

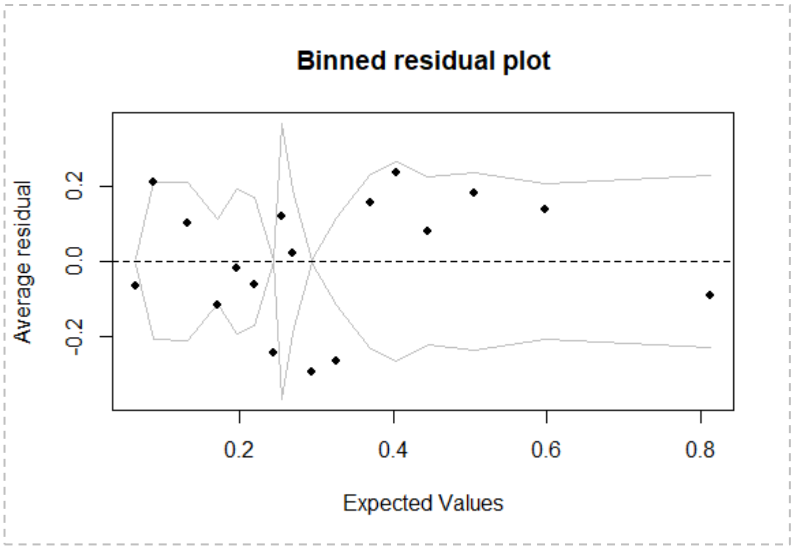


Figure 17 Binned residual plot using equation 6

This residual plot looks good and indicates a reasonable model. There are more outliers than what we would normally expect for this amount of data. In addition, we notice that there is an issue between the expected values of .2 and .3. This area of the residual plot has 3 points that are negative and lie outside of the 95% confidence interval. We decided to look deeper and see if we could find which one of the variables may have been causing this issue. Looking at the binned residual plot using the individual variable values instead of the expected values on the x-axis, we saw a similar pattern in the binned residual plot using ejection fraction. This plot is shown in figure 18. We see in figure 18 that in the middle of the data there are two points that have a negative value and lie outside the 95% confidence interval. This is like the pattern that we see in figure 17. Thus, we can conclude from these residual plots, that ejection fraction does not do a very good job predicting death from heart failure at middle range values.

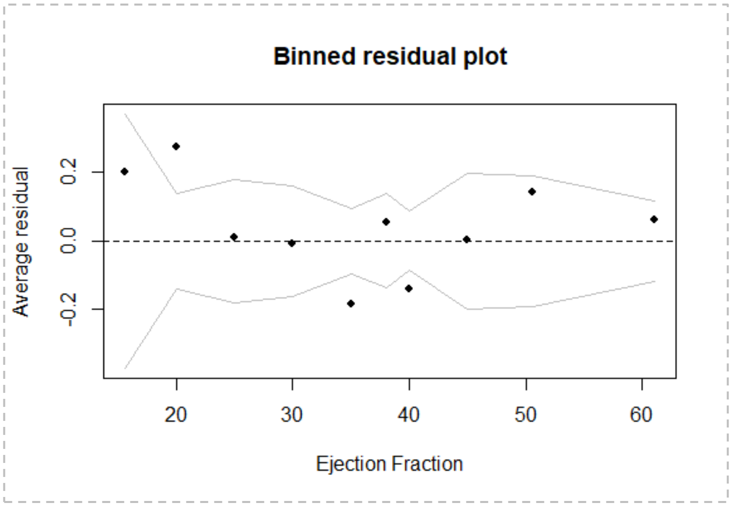


Figure 18 Binned residual plot for equation 6 residuals and ejection fraction

In addition, we made a random forest using these two variables. Table 7 shows the confusion matrix for the random forest created using equation 6.

Table 7 Confusion matrix for random forest using equation 6

	1 (patient died)	0 (patient survived)
1 (patient died)	49	47
0 (patient survived)	27	176

This random forest gave us an OOB value of 24.75%, which tells us that 75.25% of the OOB samples were correctly identified. This is better than all the previous random forests we have looked at.

We decided to also look at how all the variables affected the prediction of death from heart failure. Figure 19 shows the logistic regression curve and its corresponding probability-curve when using equation 7.

log(odds)=10.18+(-0.0021)(TIME)+(-0.534)(gender)+(-0.0135)(smoking)+.0145(diabetes)+
(-0.0103)(BP)+(-0.0075)(anemia)+0.0474(age)+(-0.0767)(ejection fraction)+
(-0.067)(sodium)+(0.666)(creatinine)+(-0.0000012)(platelets)+(0.000222)(CPK) **(7)**

The index in these graphs represent the ranking of the individual samples. They are ranked from lowest to highest in terms of probability of dying from heart failure.

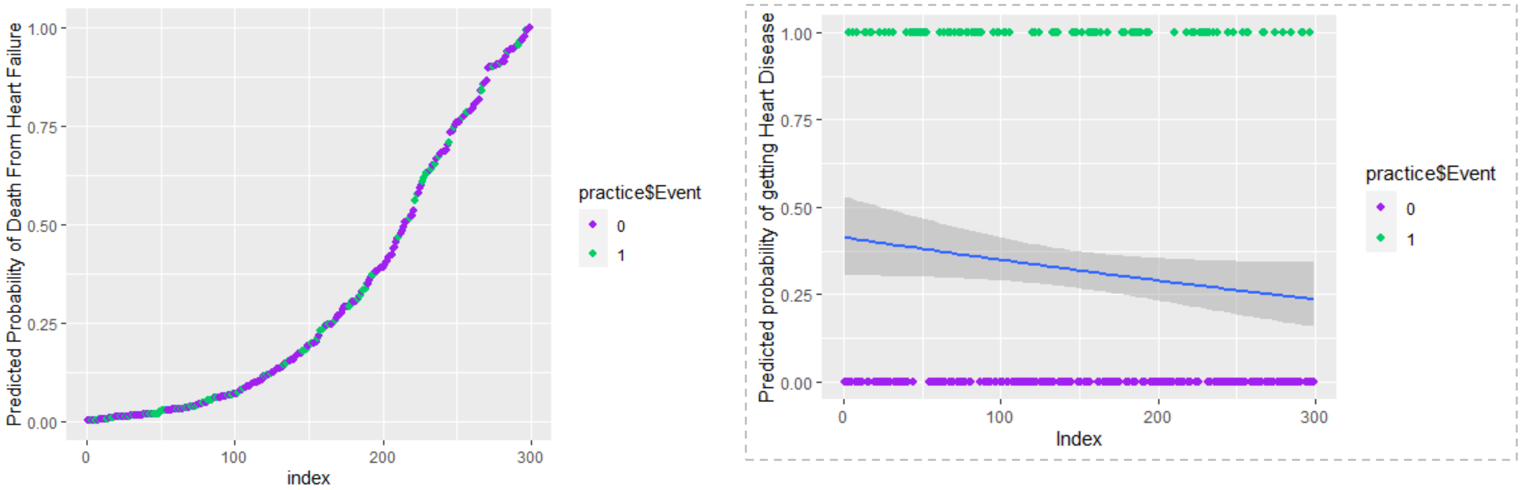


Figure 19 Logistic regression (left) and corresponding probability-curve (right) for equation 7

From this model the only variables that seemed to have a significant effect on categorizing patients were, TIME, age, ejection fraction, and creatinine, as each of these variables had a p-value that was less than 0.05. Creating a ROC graph, so that we can compare this to the other graphs we have produced, we found that this one had an AUC of 0.8972. This is close to 1, which means this is a pretty good model. As before, we used training data to look at other factors using equation 7. We got a MCC value of 0.617, which is the highest we have seen out of the four previous models. The true positive rate was 0.782, while the true negative rate was 0.862. The average accuracy was 0.839 and the average AUC was 0.822. We notice that these numbers are the highest we have seen for any of the four models thus far. However, we need to consider how much of a difference the other variables are truly making. The model that just used ejection fraction and creatinine had an accuracy of 0.749 and the AUC was 0.732. In both cases these values were increased by 0.09 when using all variables. However, the logistic regression that uses all the variables has 10 more variables than the logistic regression with just ejection fraction and creatinine. We also note that only four of the twelve variables had a significant p-value. Thus, each additional variable adds less than 0.01 to the accuracy and the AUC. This indicates that these variables are not very important and agrees with the fact that creatinine and ejection fraction predict about as good as all the variables together, as suggested in the paper [Chicco and Jurman, 2020]. We also notice that the logistic curves are similar in figures 16 and 19. The probability-curves in figure 16 and 19 are also similar. This supports the idea that the combination of creatinine and ejection fraction predicts just as well as all the variables.

Again, we looked at the binned residual plot using equation 7. This plot is given in figure 20. The grey line represents the 95% confidence interval for this data.

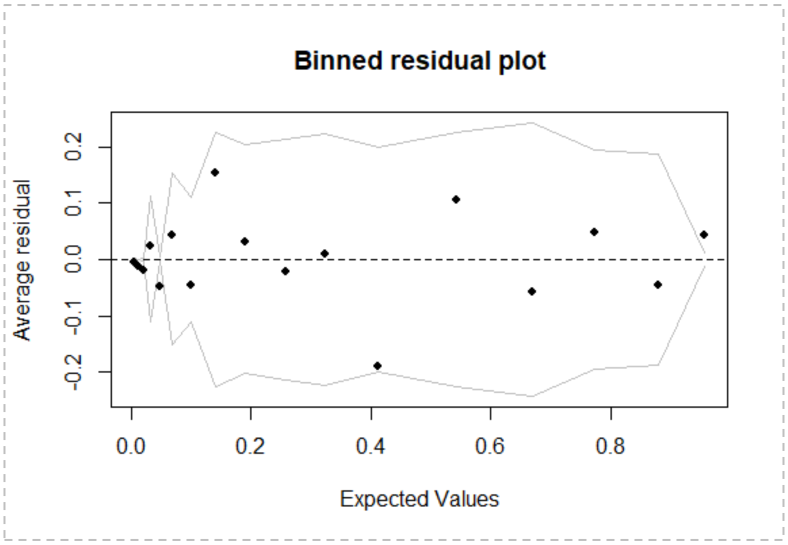


Figure 20 Binned residual plot using equation 7

This plot looks pretty good as nearly all the points lie within the 95% confidence interval. Based on this plot there does not seem to be very many outliers in this model. Out of the residual plots that we have looked at so far, this one seems to be the best. It indicates that there is not much that should be added to better predict death from heart failure. It does however ask the question if there are less variables that could be used to do just as good a job of predicting.

We again created a random forest that used all the variables. Table 8 shows the confusion matrix for this random forest.

Table 8 Confusion matrix for the random forest using equation 7

	1 (patient died)	0 (patient survived)
1 (patient died)	69	27
0 (patient survived)	17	186

This confusion matrix had an OOB value of 14.72%. This tells us that 85.28% of the OOB samples were identified correctly by this random forest. That is much higher than the other four random forests we looked at indicating that using all the variables, results in better predictions of survival.

After looking at all the information that we got from using random forests, logistic regression, and training and testing data, we determined that the random forests had the best accuracy when classifying the patients for heart disease. The highest accuracy was 85.27%, which came from the random forest created using all the possible variables. The second best accuracy came from the random forest that was created by using just ejection fraction and creatinine, with a value of 75.25%. While this accuracy is 10% less than using all the variables, we need to consider the fact that it uses 10 less variables and mathematically the accuracy must go up when more variables are added to the regression. This supports the idea that two variables do about as good as all twelve. From the residual plots that we created for the logistic regression models, we learned that all the models seem to be reliable as there were not that many major issues with any of the models. Thus, using all the variables to predict death from heart failure seems to be an inefficient way of prediction.

Appendix 4: **Various Decision Trees**

In each tree the number on the right represents the number of patients that are diagnosed with KDS for that split and the number on the left represents the number of patients that are diagnosed with ABP for that split. From this decision tree we see that more people are diagnosed with a kidney stone when they are older than 61.

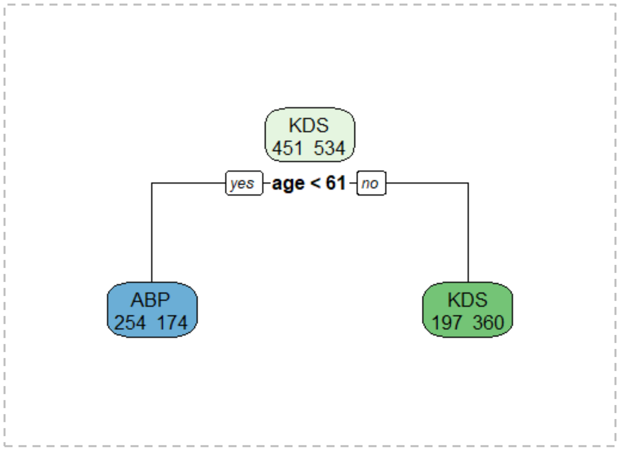


Figure 21 Decision tree using age

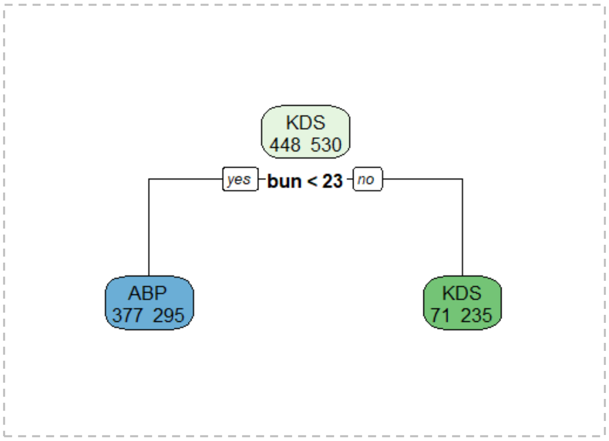


Figure 22 Decision tree using BUN

While this decision tree does not do a very good job separating the kidney stone patients, it does a pretty good job separating the patients with abdominal and back pain. We see that most of the patients that were diagnosed with abdominal and back pain had BUN levels that were lower than 23.

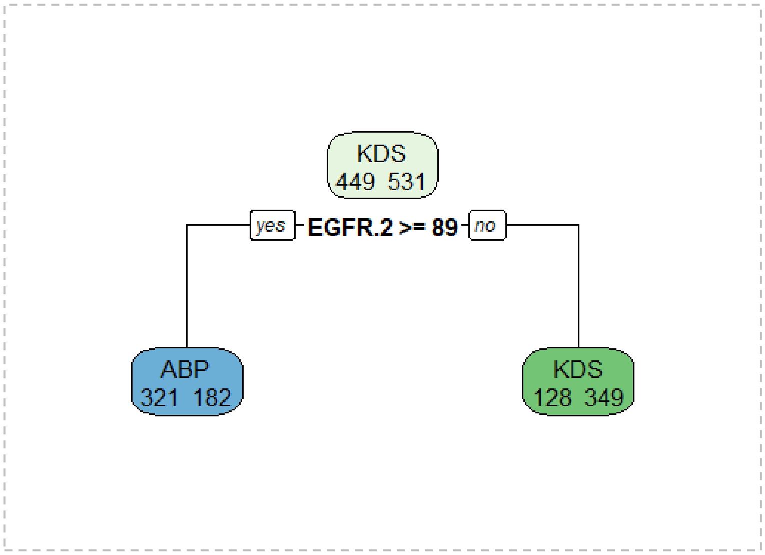


Figure 23 Decision tree using CKD-EGFR

This decision tree does a decent job of separating the two groups. Most of the patients with EGFR values less than 89 are diagnosed with kidney stones, while most of the patients with EGFR values greater than 89 are diagnosed with abdominal and back pain.

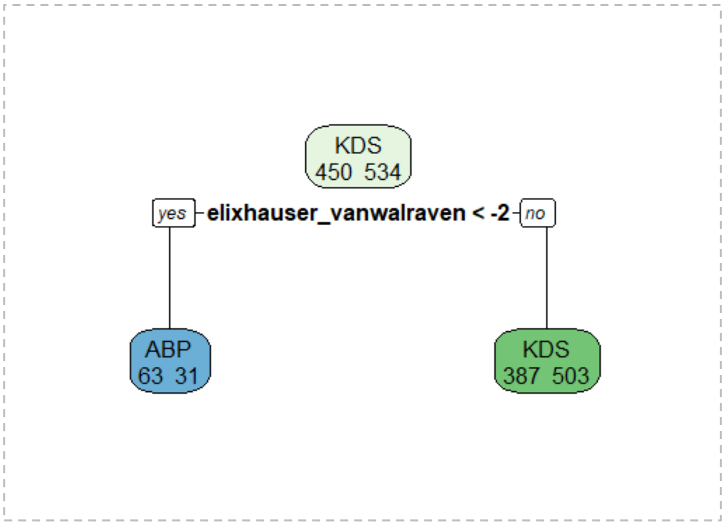


Figure 24 Decision tree using Elixhauser score

From this decision tree we see that the Elixhauser score does not do a very good job separating these two groups as nearly all the patients in the two groups have an Elixhauser score greater than -2.

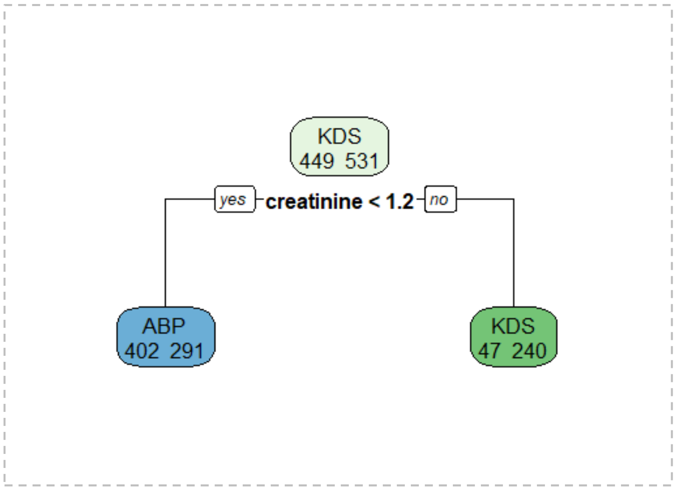


Figure 25 Decision tree using creatinine

We see that creatinine does not do a great job separating kidney stone patients as half the patients have creatinine levels less than 1.2, while the other half has creatinine values greater than 1.2. However, this seems to separate patients with abdominal and back pain well, as most of these patients have creatinine levels that are lower than 1.2.

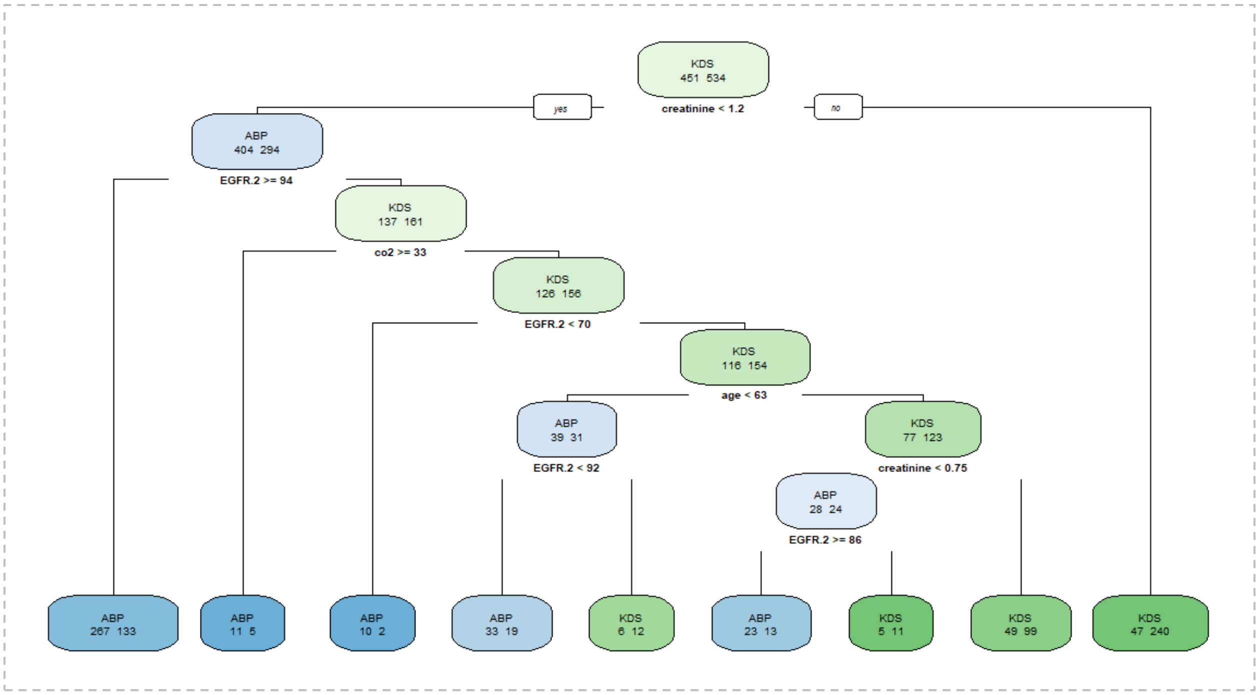


Figure 26 Decision tree using age, creatinine, BUN, CO2, and EGFR

In this decision tree, BUN was excluded from the model because it did not make any significant divisions between the two groups.

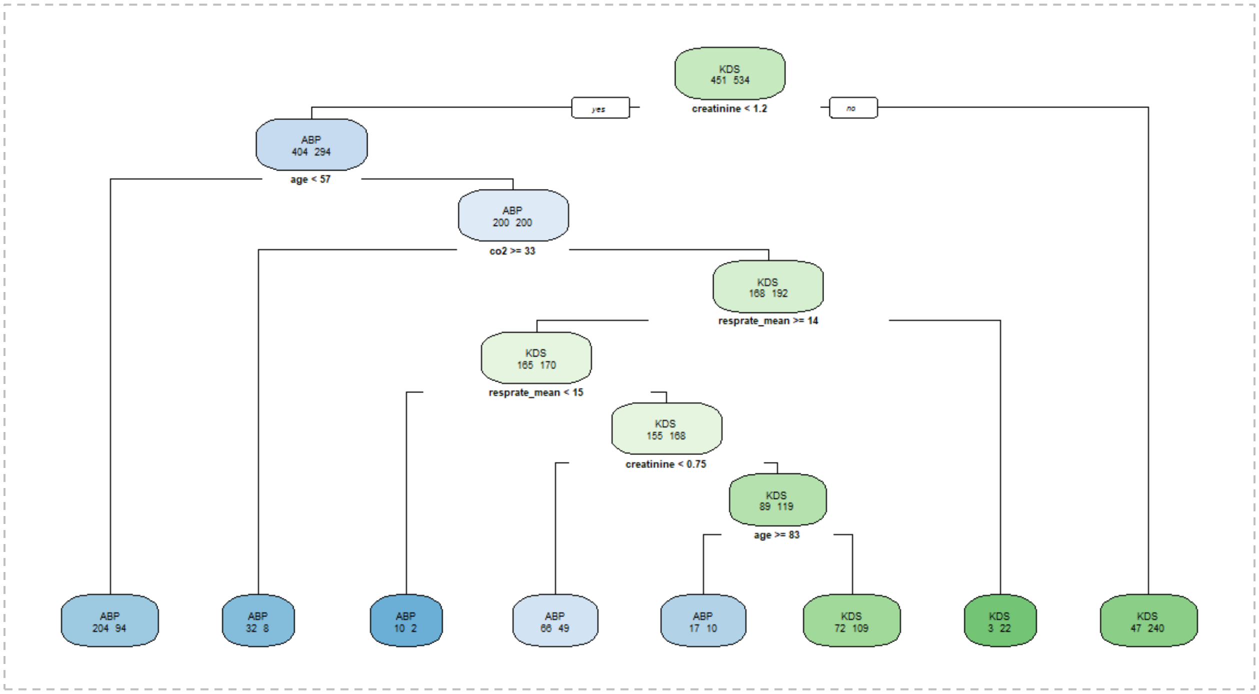


Figure 27 Decision tree using age, creatinine, chloride, CO2, and resprate mean

Appendix 5: Example Codes

```
# Remove Duplicates in Ethnicity with "other"
duplicated_id <- kds_1stlab_results[duplicated(kds_1stlab_results$hadm_id),]$subject_id
kds_1stlab_results[kds_1stlab_results$subject_id %in% duplicated_id &
kds_1stlab_results$ethnicity_grouped == "other",]
kds_1stlab_results <- kds_1stlab_results[!(kds_1stlab_results$subject_id %in% duplicated_id &
kds_1stlab_results$ethnicity_grouped == "other"),]


#Remove All Non-first Stays
kds_1stlab_results<- kds_1stlab_results[kds_1stlab_results$stay_num == "1",]

#Subset the Data to Look At Just KDS and ABP Icdtypes
kds_1stlab_results.2<-subset(kds_1stlab_results, icdtype !="OKU" & icdtype !="NCA" )
count(kds_1stlab_results$icdtype)

#Adding EGFR Calculated Using MDRD Study Equation for Use with Standardized Serum
Creatinine (Four-Variable Equation)
kds_1stlab_results<- transform(kds_1stlab_results,
EGFR=175*(kds_1stlab_results$creatinine^(-1.154))*(kds_1stlab_results$age^(-
.203))*(.742*ifelse(kds_1stlab_results$gender=="F",1,1/(.742)))*(1.210*ifelse(kds_1stlab_result
s$ethnicity_grouped=="black",1,1/(1.210))))

#Adding EGFR Using the CKD-EPI Equation for Use with Standardized Serum Creatinine
kds_1stlab_results<- transform(kds_1stlab_results,
EGFR.2=(141)*(ifelse(kds_1stlab_results$gender=="F",(kds_1stlab_results$creatinine/.7)^(-
.329),(kds_1stlab_results$creatinine/.9)^(-
.411)))*(.993^(kds_1stlab_results$age))*(1.157*(ifelse(kds_1stlab_results$ethnicity_grouped=="
black",1,1/(1.157))))*(1.018*ifelse(kds_1stlab_results$gender=="F",1,1/(1.018))))

#Logistic Regression and ROC Info Using Age
kds_1stlab_results.2<-subset(kds_1stlab_results, icdtype !="OKU" & icdtype !="NCA" )
kds_1stlab_results.2$icdtype<- ifelse(kds_1stlab_results.2$icdtype== "KDS",1,0)
logistic <- glm(icdtype~age,data=kds_1stlab_results.2,family="binomial")
summary(logistic)
library(reportROC)
reportROC(kds_1stlab_results.2$icdtype, logistic$fitted.values)

#Logistic Regression and ROC Info Using Creatinine
kds_1stlab_results.2<-subset(kds_1stlab_results, icdtype !="OKU" & icdtype !="NCA" &
creatinine !="NA" )
kds_1stlab_results.2$icdtype<- ifelse(kds_1stlab_results.2$icdtype== "KDS",1,0)
logistic <- glm(icdtype~creatinine,data=kds_1stlab_results.2,family="binomial")
summary(logistic)
reportROC(kds_1stlab_results.2$icdtype, logistic$fitted.values)
```

```
#Logistic Regression and ROC Info Using Bun
kds_1stlab_results.2<-subset(kds_1stlab_results, icdtype !="OKU" & icdtype !="NCA" & bun
!="NA" )
kds_1stlab_results.2$icdtype<- ifelse(kds_1stlab_results.2$icdtype== "KDS",1,0)
logistic <- glm(icdtype~bun,data=kds_1stlab_results.2,family="binomial")
summary(logistic)
reportROC(kds_1stlab_results.2$icdtype, logistic$fitted.values)

#Logistic Regression and ROC Info Using CDK EGFR
kds_1stlab_results.2<-subset(kds_1stlab_results, icdtype !="OKU" & icdtype !="NCA" & EGFR.2
!="NA" )
kds_1stlab_results.2$icdtype<- ifelse(kds_1stlab_results.2$icdtype== "KDS",1,0)
logistic <- glm(icdtype~EGFR.2,data=kds_1stlab_results.2,family="binomial")
summary(logistic)
reportROC(kds_1stlab_results.2$icdtype, logistic$fitted.values)

#Logistic Regression and ROC Info Using Age, Resprate Mean, Creatinine, Chloride, and
CO2
kds_1stlab_results.2<-subset(kds_1stlab_results, icdtype !="OKU" & icdtype !="NCA" & chloride
!="NA" & creatinine !="NA" & resprate_mean !="NA" & co2 !="NA" )
kds_1stlab_results.2$icdtype<- ifelse(kds_1stlab_results.2$icdtype== "KDS",1,0)
logistic <- glm(icdtype~
age+chloride+creatinine+resprate_mean+co2,data=kds_1stlab_results.2,family="binomial")
summary(logistic)
reportROC(kds_1stlab_results.2$icdtype, logistic$fitted.values)

#Logistic Regression and ROC Info Using Age, Creatinine, Bun, CKD-EGFR and CO2
kds_1stlab_results.2<-subset(kds_1stlab_results, icdtype !="OKU" & icdtype !="NCA" & bun
!="NA" & creatinine !="NA" & EGFR.2 !="NA" & co2 !="NA" )
kds_1stlab_results.2$icdtype<- ifelse(kds_1stlab_results.2$icdtype== "KDS",1,0)
logistic <- glm(icdtype~
age+bun+creatinine+EGFR.2+co2,data=kds_1stlab_results.2,family="binomial")
summary(logistic)
reportROC(kds_1stlab_results.2$icdtype, logistic$fitted.values)

#Logistic Regression and ROC Info Using Elixhauser Score
kds_1stlab_results.2<-subset(kds_1stlab_results, icdtype !="OKU" & icdtype !="NCA" &
elixhauser_vanwalraven !="NA" )
kds_1stlab_results.2$icdtype<- ifelse(kds_1stlab_results.2$icdtype== "KDS",1,0)
logistic <-glm(icdtype~elixhauser_vanwalraven,data=kds_1stlab_results.2,family="binomial")
summary(logistic)
reportROC(kds_1stlab_results.2$icdtype, logistic$fitted.values)
```

#Statistical Summaries

```
Library(plyr)
summary(kds_1stlab_results.2$age)
summary(kds_1stlab_results.2$EGFR.2)
count(kds_1stlab_results.2$ethnicity_grouped)
count(kds_1stlab_results.2$gender)
summary(kds_1stlab_results.2$creatinine)
summary(kds_1stlab_results.2$bmi)
summary(kds_1stlab_results.2$elixhauser_vanwalraven)
```

#A few Correlations

```
cor(kds_1stlab_results.2$heartrate_mean,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$sysbp_mean,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$diasbp_mean,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$meanbp_mean,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$resprate_mean,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$tempc_mean,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$spo2_mean,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$glucose_mean,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$anion_gap,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$albumin,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$bands,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$base_excess,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$bicarbonate,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$bilirubin,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$calcium,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$creatine_kinase,kds_1stlab_results.2$icdtype,use = "complete.obs")
cor(kds_1stlab_results.2$creatinine,kds_1stlab_results.2$icdtype,use = "complete.obs")
```

#Categorizing Healthy vs Not Healthy (just a few we did)
#Cleaning and Combining Data

```
kds_oth<- transform(kds_oth,
EGFR.2=(141)*(ifelse(kds_oth$gender=="F",(kds_oth$creatinine/.7)^(-
.329),(kds_oth$creatinine/.9)^(-
.411)))*(.993^(kds_oth$age))*(1.157*(ifelse(kds_oth$ethnicity_grouped=="black",1,1/(1.157))))*(
1.018*ifelse(kds_oth$gender=="F",1,1/(1.018))))

 kds_4<-rbind(kds_oth,kds_1stlab_results)

duplicated_id <- kds_4[duplicated(kds_4$hadm_id),]$subject_id
kds_4[kds_4$subject_id %in% duplicated_id & kds_4$ethnicity_grouped == "other",]
kds_4 <- kds_4[!(kds_4$subject_id %in% duplicated_id & kds_4$ethnicity_grouped == "other"),]
```

```
#Creatinine
kds_4.2<-subset(kds_4, creatinine !="NA" )
kds_4.2$creatininerange <- findInterval(kds_4.2$creatinine, c(.6,1.2))
kds_4.2$creatininerange <- as.factor(kds_4.2$creatininerange)
levels(kds_4.2$creatininerange) <- c("not healthy -","healthy","not healthy +")
library(dplyr)
kds_4.2 %>% group_by(kds_4.2$icdtype,kds_4.2$creatininerange) %>%
summarise(count_sales = n())
```

```
# pH
kds_4.2<-subset(kds_4, ph !="NA" )
kds_4.2$phrange <- findInterval(kds_4.2$ph, c(7.35,7.45))
kds_4.2$phrange <- as.factor(kds_4.2$phrange)
levels(kds_4.2$phrange) <- c("not healthy -","healthy","not healthy +")
kds_4.2 %>% group_by(kds_4.2$icdtype,kds_4.2$phrange) %>% summarise(count_sales =
n())
```

```
#Chloride
kds_4.2<-subset(kds_4, chloride !="NA" )
kds_4.2$chloriderange <- findInterval(kds_4.2$chloride, c(98,106))
kds_4.2$chloriderange <- as.factor(kds_4.2$chloriderange)
levels(kds_4.2$chloriderange) <- c("not healthy -","healthy","not healthy +")
kds_4.2 %>% group_by(kds_4.2$icdtype,kds_4.2$chloriderange) %>% summarise(count_sales
= n())
```

```
#Average Data Random Forests, Regression Curves, and Residual Plots. (KDS VS NCA)
#Cleaning the Data
duplicated_id <- kds_data2[duplicated(kds_data2$hadm_id),]$subject_id
kds_data2[kds_data2$subject_id %in% duplicated_id & kds_data2$ethnicity_grouped ==
"other",]
kds_data2 <- kds_data2[!(kds_data2$subject_id %in% duplicated_id &
kds_data2$ethnicity_grouped == "other"),]
```

```
#Subset the Data
kds_data2.2<-subset(kds_data2, icdtype !="OKU" & icdtype !="NCA" )
```

```
#Logistic Regression Using Age
kds_data2.2$icdtype<- ifelse(kds_data2.2$icdtype== "KDS",1,0)
logistic <- glm(icdtype~age,data=kds_data2.2,family="binomial")
summary(logistic)
```

```
#Logistic Regression Curve
kds_data2.2$icdtype<- as.factor(kds_data2.2$icdtype)
```

```
predicted.data <- data.frame(probability.of.KDS=logistic$fitted.values,
KDS=kds_data2.2$icdtype)
ggplot(data=predicted.data, aes(x=kds_data2.2$age, y=ifelse(kds_data2.2$icdtype == "1", 1,
0))) +
geom_point(aes(color=kds_data2.2$icdtype), alpha=1, stroke=1) +
geom_smooth(method = "glm", method.args = list(family = "binomial")) +
xlab("Index") +
ylab("Predicted probability of getting Kidney Stones")+scale_color_manual(values = colors)

#Residual Plot
Library("arm")
binnedplot(fitted(logistic),
residuals(logistic, type = "response"),
nclass = NULL,
xlab = "Expected Values",
ylab = "Average residual",
main = "Binned residual plot",
cex.pts = 0.8,
col.pts = 1,
col.int = "gray")

#ROC Curve
library(pROC)
par(pty="s")
roc(kds_data2.2$icdtype, logistic$fitted.values, plot=TRUE, legacy.axes=TRUE)

#Random Forests
Library(randomForest)
kds_data2.2<-subset(kds_data2, icdtype !="OKU" & icdtype !="NCA" )
kds_data2.2$icdtype<- ifelse(kds_data2.2$icdtype== "KDS",1,0)
logistic <- glm(icdtype~age,data=kds_data2.2,family="binomial")
summary(logistic)
kds_data2.2$age<-as.numeric(kds_data2.2$age)
kds_data2.2$icdtype<-as.factor(kds_data2.2$icdtype)
model <- randomForest(icdtype~age, data=kds_data2.2, proximity=TRUE)
Model

#Feature Importance Plot
kds_1stlab_results.2<-subset(kds_1stlab_results, icdtype !="NCA" & icdtype != "OKU" & bun
!="NA" & elixhauser_vanwalraven !="NA"& EGFR.2 !="NA" & bands !="NA" & chloride !="NA" &
creatinine !="NA" & resprate_mean !="NA" & co2 !="NA" )
kds_1stlab_results.2$icdtype<- ifelse(kds_1stlab_results.2$icdtype== "KDS",1,0)
```

```
model_rf <- ranger(icdtype ~
age+bun+elixhauser_vanwalraven+EGFR.2+creatinine+resprate_mean+co2+chloride+bands,d
ata = kds_1stlab_results.2 ,importance = "impurity")
argIDCART = rpart(icdtype ~
age+bun+elixhauser_vanwalraven+EGFR.2+creatinine+resprate_mean+co2+chloride+bands,
data = kds_1stlab_results.2,   method = "class")
argPlot <- as.data.frame(argIDCART$variable.importance)
argPlot
library(rpart)
library(tidyverse)
fit <- rpart(icdtype ~
age+bun+elixhauser_vanwalraven+EGFR.2+creatinine+resprate_mean+co2+chloride+bands,
data = kds_1stlab_results.2)
df <- data.frame(imp = argIDCART$variable.importance)
df2 <- df %>%
tibble::rownames_to_column() %>%
dplyr::rename("variable" = rowname) %>%
dplyr::arrange(imp) %>%
dplyr::mutate(variable = forcats::fct_inorder(variable))
ggplot2::ggplot(df2) +
geom_col(aes(x = variable, y = imp),
col = "black", show.legend = F) +
coord_flip() +
scale_fill_grey() +
theme_bw()
```

References

[eGFR, 2020] About eGFR. (2020). Retrieved from https://renal.org/information-resources/the-uk-eCKD-guide/about-egfr/

[Agrawal, 2020] Agrawal, N. (2020, June 02). What is Chi-Square Test? Chi- Square Test Explained. Retrieved from https://www.mygreatlearning.com/blog/chi-square-test-explained/

[Bloodbook, 2013] “BLOOD TEST RESULTS - NORMAL RANGES.” Blood Test Results with Normal Range Reference Chart - BloodBook, Blood Information for Life, Bloodbook.com, 2013, bloodbook.com/ranges.html.

[BMI] BMI weight categories. (n.d.). Retrieved from https://www.medic8.com/nutrition/bmi/weight-categories.html

[Bradburn, 2020] Bradburn, S. (2020, June 08). What Is Pearson Correlation? Including Test Assumptions. Retrieved from https://toptipbio.com/what-is-pearson-correlation/

[Chen, et.al., 2018] Chen, Z., Bird, V. Y., Ruchi, R., Segal, M. S., Bian, J., Khan, S. R., Elie, M. C., & Prosperi, M. (2018). Development of a personalized diagnostic model for kidney stone disease tailored to acute care by integrating large clinical, demographics and laboratory data: the diagnostic acute care algorithm - kidney stones (DACA-KS). *BMC medical informatics and decision making*, *18*(1), 72. https://doi.org/10.1186/s12911-018-0652-4

[Chicco and Jurman, 2020] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

[Ciorniciuc, 2015] Ciorniciuc, V. (2015, July 25). Charlson Comorbidity Index (CCI) Calculator. Retrieved from https://www.thecalculator.co/health/Charlson-Comorbidity-Index-(CCI)-Calculator-765.html

[Comparing Frequencies] Comparing Frequencies. (2018, October 18). Retrieved from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717_ComparingFrequencies/PH717_ComparingFrequencies8.html

[Daniels, et. al., 2016] Daniels, B., Gross, C. P., Molinaro, A., Singh, D., Luty, S., Jessey, R., & Moore, C. L. (2016). STONE PLUS: Evaluation of Emergency Department Patients With Suspected Renal Colic, Using a Clinical Prediction Tool Combined With Point-of-Care Limited Ultrasonography. *Annals of emergency medicine*, *67*(4), 439–448. https://doi.org/10.1016/j.annemergmed.2015.10.020

[Donges, 2019] Donges, Niklas. (2019). A complete guide to the random forest algorithm. Retrieved from https://builtin.com/data-science/random-forest-algorithm

[Fortney, 2018] Fortney, K. (2018, January 10). Machine Learning - An Error by Any Other Name... Retrieved from https://towardsdatascience.com/machine-learning-an-error-by-any-other-name-a7760a702c4d

[ECI, 2021] Free Online Elixhauser Comorbidity Index Calculator. (2021). Retrieved from https://orthotoolkit.com/elixhauser-comorbidity-index/

[Gilbert & Weiner, 2014] Gilbert, S. J., & Weiner, D. E. (2014). *NATIONAL KIDNEY FOUNDATION’S PRIMER ON KIDNEY DISEASES* (6th ed.) (D. S. Gipson, M. A. Perazella, & M. Tonelli, Eds.). Philadelphia, PA: National Kidney Foundation.

[Grace-Martin, 2018] Grace-Martin, K. (2018, December 13). What Is an ROC Curve? Retrieved from https://www.theanalysisfactor.com/what-is-an-roc-curve/

[Kidney Stones, (a)] Kidney stones. (2020, May 05). Retrieved from https://www.mayoclinic.org/diseases-conditions/kidney-stones/symptoms-causes/syc-20355755.

[Kidney Stones, (b)] Kidney stones - Genetics Home Reference - NIH. (2020, August 17). Retrieved from https://ghr.nlm.nih.gov/condition/kidney-stones#resources

[Kidney Stones, (c)] Kidney Stones: Symptoms, Causes, Diagnosis, Treatment & Prevention. (2020). Retrieved September 28, 2020, from https://my.clevelandclinic.org/health/diseases/15604-kidney-stones.

[IQWiG] “Kidney Stones: Overview.” InformedHealth.org [Internet]., Institute for Quality And Efficiency in Health Care (IQWiG), 28 Feb. 2019, www.ncbi.nlm.nih.gov/books/NBK348937/.

[Jiang, et. al., 2019] Jiang, K., Sun, F., Zhu, J. *et al.* Evaluation of three stone-scoring systems for predicting SFR and complications after percutaneous nephrolithotomy: a systematic review and meta-analysis. *BMC Urol* 19, 57 (2019). https://doi.org/10.1186/s12894-019-0488-y

[Khan, 2016] Khan, S. R., Pearle, M. S., Robertson, W. G., Gambaro, G., Canales, B. K., Doizi, S., Traxer, O., & Tiselius, H. G. (2016). Kidney stones. *Nature reviews. Disease primers*, *2*, 16008. https://doi.org/10.1038/nrdp.2016.8

[Krishni, 2019] Krishni. (2019, June 05). A Beginners Guide to Random Forest Regression. Retrieved from https://medium.com/datadriveninvestor/random-forest-regression-9871bc9a25eb

[Livingston, 2005] Livingston, F. (2005). *Implementing Breiman’s Random Forest Algorithm into Weka*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.607.8926&rep=rep1&type=pdf

[Logistic Regression] Lesson 6: Logistic Regression. (2018). Retrieved from https://online.stat.psu.edu/stat504/node/149/

[Mankad, 2019] Mankad, R. (2019, July 02). Ejection fraction: What does it measure? Retrieved from https://www.mayoclinic.org/ejection-fraction/expert-answers/FAQ-20058286

[May, et.al., 2019] May, P. M., Rowland, D., & Murry, K. (2019). JESS - JOINT EXPERT SPECIATION SYSTEM. Retrieved from http://jess.murdoch.edu.au/docs/Jess_Primer_V86.pdf

[MIMIC-III] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635

[Mokobi, 2020] Mokobi, F., & Gayyas, M. (2020, April 18). What is Sensitivity, Specificity, False positive, False negative? Retrieved from https://microbenotes.com/sensitivity-specificity-false-positive-false-negative/

[Moore, 2020] Moore, C. L. (2020). STONE Score for Uncomplicated Ureteral Stone. Retrieved from https://www.mdcalc.com/stone-score-uncomplicated-ureteral-stone#why-use

[Myers, et. al., 2018] Myers, J., Goodenow, D., Gokoglu, S., & Kassemi, M. (january 22, 2018).

Sensitivity Analysis of the Change of Renal Stone Occurrence Rates in Astronauts Using Urine Chemistries [Presentation]. Galveston, TX.

[Nall,2021] Nall, R. (2021). Glomerular Filtration Rate Test. Retrieved from https://www.healthline.com/health/glomerular-filtration-rate

[Narkhede, 2019] Narkhede, S. (2019, May 26). Understanding AUC - ROC Curve. Retrieved from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[Nedea, 2017] Nedea, D. (2017, May 30). Kidney STONE Score Calculator. Retrieved from https://www.mdapp.co/kidney-stone-score-calculator-263/

[Odds Ratio] Odds ratio - Confidence Interval. (2020, April 09). Retrieved from https://select-statistics.co.uk/calculators/confidence-interval-calculator-odds-ratio/

[R Tutorial, 2020] R Tutorial for Beginners: Learn R Programming Language. (2020). Retrieved from https://www.guru99.com/r-tutorial.html

[Rodgers, et. al., 2014] Rodgers A, Gauvin D, Edeh S, Allie-Hamdulay S, Jackson G, et al. (2014) Sulfate but Not Thiosulfate Reduces Calculated and Measured Urinary Ionized Calcium and Supersaturation: Implications for the Treatment of Calcium Renal Stones. PLoS ONE 9(7): e103602. doi:10.1371/journal.pone.0103602

[Rouse, 2018] Rouse, M. (2018, April 18). What is confusion matrix? - Definition from WhatIs.com. Retrieved from https://whatis.techtarget.com/definition/confusion-matrix

[Safaie, et. al., 2019] Safaie, A, Mirzadeh, M, Aliniagerdroudbari, E, et al. A clinical prediction rule for uncomplicated ureteral stone: the STONE score; a prospective observational validation cohort study. Turk J Emerg Med 2019; 19(3): 91–95.

[SPSS, 2020] SPSS Tutorials: Pearson Correlation. (2020). Retrieved from https://libguides.library.kent.edu/SPSS/PearsonCorr

[Starmer] Starmer, J. StatQuest videos. Retrieved October 10, 2020, from https://www.youtube.com/c/joshstarmer

[Stephens, 2020] Stephens, C. (2020). Hydronephrosis. Retrieved from https://www.healthline.com/health/unilateral-hydronephrosis

[Webb,2017] Webb, J. (2017, September 03). Course Notes for IS 6489, Statistics and Predictive

Analytics. Retrieved from https://bookdown.org/jefftemplewebb/IS-6489/logistic-regression.html#assessing-logistic-model-fit

[Supersaturated, 2020] What Is a Supersaturated Solution? (2020). Retrieved from https://www.reference.com/science/supersaturated-solution-a40ff09bc3e97f24

[R project] The R Project for Statistical Computing. (n.d.). Retrieved from http://www.r-project.org/

[Decision Tree] “What Is Decision Tree?: Comprehensive Guide to Decision Tree.” *EDUCBA*, 17 Apr. 2020, www.educba.com/what-is-decision-tree/.

[Zach, 2020] Zach. (2020, March 02). How to Interpret Odds Ratios. Retrieved from https://www.statology.org/interpret-odds-ratios/