STATISTICS & DATA SCIENCE SENIOR PROJECT

Predicting Acute Kidney Injury After Percutaneous Coronary Intervention Using Longitudinal Lab Data

Imad Rizvi[1] | **Chenxi Huang (Advisor)**[2]

[1]Yale University, New Haven, CT, 06520, USA

[2]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, 06510, USA

Correspondence
Email: imad.rizvi@yale.edu

Acute kidney injury (AKI) is a frequent complication that occurs after percutaneous coronary intervention (PCI) in 7-10% of cases. Several studies over the past decade have worked on models to predict incidence of AKI after PCI from a variety of risk factors to inform decisions on treatment and mitigation strategies. Using a dataset of electronic health records (EHR) from Yale New Haven Hospital (YNHH), we explored the application and development of advanced machine learning tools in extracting previously unrecognized patterns in longitudinal lab measurements to enhance the prediction performance of post-PCI AKI models. We include variables for patient demographics, prior procedures, prior medications, prior diagnoses, and most recent lab value measurements as static variables. We use a baseline logistic regression model, various feedforward neural networks, and LSTM models. We find that including multiple (sequential) pre-procedure creatinine measures improves the performance over static variables. Additionally, we find that our current LSTM architecture does not improve upon the feedforward neural network for modeling the longitudinal lab data, though future work should explore if different architectures or methods of including the irregularly-sampled, sequential pre-procedure creatinine values can lead to improved performance.

1 | INTRODUCTION

Coronary artery disease is the leading cause of mortality in the United States accounting for over 600,000 annual deaths [1]. Acute coronary syndrome is a subcategory of coronary artery disease [2], and a common and effective treatment option is percutaneous coronary intervention (PCI) [3]. In order to perform PCI, a contrast dye is injected for diagnostic imaging, and due to the potential nephrotoxicity of iodine contrast agents, acute kidney injury (AKI) is one of the most common complications after PCI, increasing patient morbidity [3]. Depending on the definition used for AKI, which is usually measured by increase in creatinine level, the incidence of AKI after PCI can range from from around 7-10% [4][5]. Thus, developing a better understanding of various factors that may predict AKI after PCI is crucial for reducing patient morbidity, as varying amounts of contrast can be used or the procedure can be avoided entirely.

Over the past few years, several studies have either validated existing models for predicting AKI or developed new and improved models. In 2018, *Huang et. al.* aimed to apply various machine learning techniques to improve upon a contemporary logistic regression model [4]. The authors found a small but statistically significant improvement in performance over the baseline AKI model using the same dataset and found that the new model performed better for patients with extremely low and high risks [4]. More recently, in 2022, *Kuno et. al.* applied a tree-based ML approach aiming to predict AKI with a lightGBM model, and the authors noted that their model was able to quantify the risk of AKI with almost half the number of clinical variables compared to a baseline logistic regression risk model [5]. Despite the several published models, these studies have largely relied on cross-sectional data and ignored the potential predictive information contained in the longitudinal vital and lab measurements.

Using electronic health records (EHR), which contain richer data and entire care history of patients, we aim to explore the application and development of advanced machine learning tools in extracting previously unrecognized patterns to enhance the prediction performance of post-PCI AKI models.

2 | METHODOLOGY

2.1 | Dataset

We use EHR data from patients receiving PCI at Yale New Haven Hospital (YNHH). The dataset consists of basic demographics, prior health information, and longitudinal lab values from patients up to 1 year before the procedure, though the frequency, quantity, and dates of measurements are not consistent across patients. The demographic information includes reported race, ethnicity, age, and sex. Additionally, we had access to data on prior patient diagnoses, medication histories, prior admits (whether they were emergency or elective), and prior imaging procedures. Finally, we had access to longitudinal lab values of creatinine, along with measurements of sodium, chloride, blood urea nitrogen (BUN), potassium, hemoglobin, and other substances. The original dataset consisted of 11,662 unique PCI procedures from 9,386 unique patients. These procedures took place between January 2013 and November 2019.

2.2 | Preprocessing

The initial exclusion criteria were in line with prior literature [4][6] and are summarized in Figure 1.

These exclusion criteria consisted of removing procedures where patients were discharged on the same day, procedures that were not the first in a single hospitalization, procedures without a before or after creatinine measurement, and procedures where the patient was on dialysis. Overall, this narrowed the dataset from 11,662 procedures from 9,386 unique patients to 7,146 procedures from 6,146 unique patients.

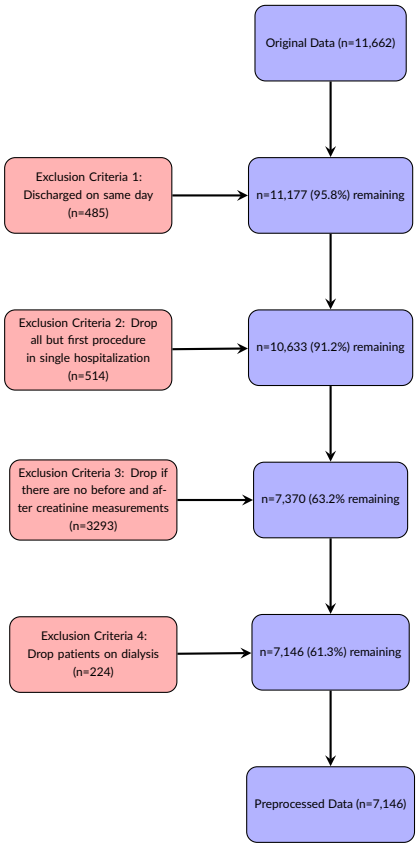

FIGURE 1 The exclusion criteria applied in order. The "n" number represents the number of procedures at each step

After implementing the exclusion criteria, we next needed to define our outcome variable – the incidence of AKI. This was computed by using a standard definition of AKI as a >= .3 mg/dl absolute increase in creatinine levels or a 1.5x relative increase in creatinine levels [6][4]. The prevalence of AKI in the dataset was 11.7%.

Within the remaining procedures after applying the exclusion criteria, 70.9% involved male patients and 29.1% involved female patients. The patients had a mean age of 80.0 with a standard deviation of 12.1. The primary race reported was 82.2% White or Caucasian, 8.6% Black or African American, 7.5% other/refused/unknown, 1.4% Asian, .1% Native Hawaiian or Other Pacific Islander, and .1% American Indian or Alaska Native.

2.3 | Feature Engineering

Feature engineering was required to transform the basic demographics, prior health information, and longitudinal lab values into features that could be inputted into various models. All categorical demographic information (including primary race, ethnicity and sex) was one-hot encoded. For some of the pre-procedure measurements (height, weight, diastolic blood pressure, systolic blood pressure, temperature, pulse, respiration), data was missing for less than 5% of procedures total across all these variables. Thus, missing data was imputed using the mean across all patients, and indicator variables were created to indicate whether the value was imputed (only four were necessary since the blood pressure, temperature, and pulse measurements had missing data in the same rows). For lab values, we included a feature for any labs for which we had data prior to procedure in at least 90% of the data rows (there were 10 such values including creatinine, the 11th highest was only appearing in 45.6% of the rows). For cases where there was missing data, we imputed the data using the mean lab value over all patients, and created a separate indicator feature to mark where values had been imputed.

For all the following medical history-related variables, we only included any information that occurred prior to hospital admission for the PCI procedure. For prior medication history, we one-hot encoded pharmaceutical classes and therapeutic classes of drugs that were prescribed in more than 25% of the cases in the data (this included 31 pharmaceutical classes of drugs and 24 therapeutic classes of drugs. For prior diagnosis history, we one-hot encoded any prior diagnoses that were diagnosed before admission to the hospital for the procedure if they appeared in more than 25% of the cases in the data. These were based on ICD10 codes and the most common diagnoses were hypertension (73.3%), hyperlipidemia (62.7%), and atherosclerotic heart disease of native coronary artery without angina pectoris (60.5%). In total, 11 prior diagnoses were one-hot encoded. For prior imaging procedures, since these were more sparse, we one-hot encoded any imaging procedures if they occurred in more than 10% of the cases, and there were 9 such imaging procedures.

At the end of this feature engineering, we had 115 static predictors (no timing information had been encoded) to use to predict our outcome variable (AKI) in the baseline model.

For the dynamic data, we marked all prior creatinine measurements along with the date and time they were taken. We also include dynamic data for all the 9 other labs we had previously included because we had found we had measurements for at least 90% of the rows. These included blood urea nitrogen, chloride, sodium, potassium, hematocrit, hemoglobin, wbc, neutrophil absolute count, and neutrophils. In order to be able to incorporate this data into a feed-forward neural network, we created various buckets (date ranges pre-procedure) that balanced the numbers of observations in each bucket (Figure 2). The buckets were (1) up to 48 hours pre-procedure, (2) 48 hours - 1 month pre-procedure, (3) 1 month - 6 months pre-procedure, and (4) 6 months - 1 year pre-procedure. Figure 2 shows the range of number of measurements in each bucket.

Table 1 shows a summary of features included in each of the models.

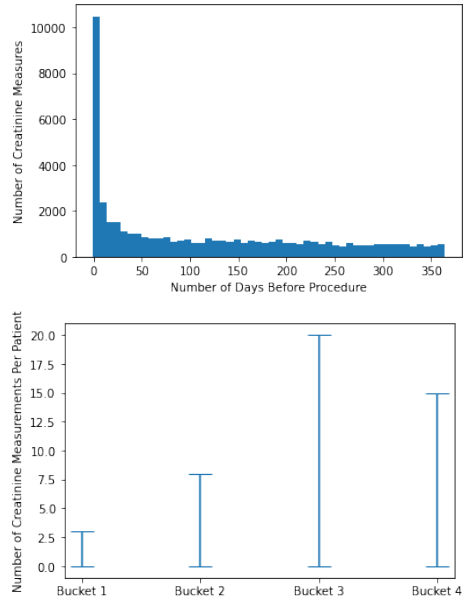

FIGURE 2 The above plot (histogram) shows all creatinine measures, and how many days before the patient's respective procedure they were taken. The below plot shows for each bucket the range (excluding the 5% most extreme values) of number of creatinine measurements per patient.

3 | RESULTS

3.1 | Baseline Logistic Regression

Before attempting the neural network models, we first sought to use our 115 static predictors in a logistic regression model to determine a baseline performance. We performed 5-fold cross validation with random splits, though we ensured that no patient appeared in multiple folds to avoid data

Model	# Features	Static Features Included	Dynamic Features Included
Baseline (Section 3.1/3.2)	115	Demographics, Most Recent Lab Values, Medication History, Diagnosis History, Procedure History	-
Creatinine Buckets (Section 3.3.1)	123	Demographics, Most Recent Lab Values, Medication History, Diagnosis History, Procedure History	4 Creatinine Buckets
Changing Creatinine Values (Section 3.3.2)	132	Demographics, Most Recent Lab Values, Medication History, Diagnosis History, Procedure History	4 Creatinine Buckets, 3 Differences (min, mean, max) Between Sequential Buckets
All Changing Lab Values (Section 3.3.3)	285	Demographics, Most Recent Lab Values, Medication History, Diagnosis History, Procedure History	4 Buckets + 3 Differences (min, mean, max) Between Sequential Buckets for 10 Lab Values

TABLE 1 This table shows the different features in each of the models.

leakage. We repeated this method 20 times to generate a confidence interval for the model performance. We found that the mean value of the area under the ROC Curve (AUC) for the test data was .806. We found that the 95% confidence interval for the model performance for this metric was .805-.812.

3.2 | Baseline Neural Network Model

Next, we used the same data on a neural network model to understand the baseline neural network model performance without any of the dynamic variables. We performed nested cross validation with five folds in both the outer and inner cross validation loops in order to tune neural network parameters including batch size, width of intermediate layers, learning rate, and number of epochs for training. Our neural network consisted of one middle layer with its width tuned as a hyperparameter. A tanh activation function was applied to the middle layer and a sigmoid activation function was applied to the output layer. We used an Adam optimizer and binary cross entropy loss function.

We found that the average AUC score for the test data was .810, and the 95% confidence interval was .807-.814.

We also experimented with deeper neural networks, such as a neural network with two middle layers and a neural network with three middle layers. We found that these deeper neural network models had average test data AUC scores of .808 (.805-.812 95% CI) and .807 (.803-.811 95% CI) respectively. Thus, since there was not a significant increase in model performance by incorporating additional layers, we retained the same architecture with one intermediate layer for other models as well.

3.3 | Incorporating Longitudinal Lab Values

3.3.1 | Incorporating Creatinine Measurement Buckets

Next, we incorporated the longitudinal creatinine lab values. As described in Section 2, we had creatinine measurements from up to 1 year pre-procedure for patients. We placed all these measurements into four different buckets based on time before procedure, and these buckets were chosen in order to reduce the number of instances of empty buckets for a patient (see Section 2). For each bucket, we had a variable for the mean measure, along with an indicator variable to represent whether or not the data had to be imputed. We imputed values using the mean value of that bucket over all patients in the data. We then repeated the same nested cross validation process with the same neural network architecture as specified in the previous section. We found that the average AUC score for the test data was .819, and the 95% confidence interval was .815-.823. Thus, incorporating these buckets compared to just incorporating a single pre-procedure creatinine measure led to a significant improvement in model performance.

3.3.2 | Incorporating Changing Creatinine Values

In order to capture the dynamics of changing creatinine values over times, we then created different features out of the four buckets, including average sequential difference, minimum sequential difference (smallest in the earlier period minus largest in the later period), and maximum sequential difference (largest in the earlier period minus smallest in the later period). We used the same neural network architecture as previously specified. We found that the average AUC score for the test data was .818, and the 95% confidence interval was .815-.822. Thus, including these additional features had no additional explanatory power for our model.

3.3.3 | Incorporating All Changing Lab Values

Finally, we incorporated all the lab values for which we had sufficient measures, rather than just creatinine. These were the same substances as specified in Section 2, and included blood urea nitrogen, chloride, sodium, potassium, hematocrit, hemoglobin, wbc, neutrophil absolute count, and neutrophils. We created the same four buckets as we had for the creatinine measures, and took the same sequential differences as specified above (minimum difference, maximum difference, mean difference). We used the same neural network architecture as previously specified, and we found that the average AUC score for the test data was .818, and the 95% confidence interval was .814-.821. We found no evidence that including sequential data for lab values besides creatinine led to improvements in our neural network model performance.

3.3.4 | LSTM Dynamic Model

LSTM models are a type of recurrent neural network that is capable of learning dependencies in sequential data. These models have been used in other areas of biological research [7][8]. Thus, we aimed to see if using an LSTM model could improve our performance for the sequential data. The neural network architecture that was used is specified in Figure 3. We created an LSTM layer that took in our 4 time-steps (one for each bucket) for any dynamic features (creatinine or other lab values). The LSTM layer consisted of 20 hidden states. We then combined the output of this LSTM layer with the static predictors we used in our other models, and created a fully connected layer with a width of 60, applying a tanh activation function. Finally, this layer was connected to the output layer and a sigmoid activation function was applied to generate our predictions for risk of AKI after PCI.

3.3.5 | LSTM With Creatinine Values

We first found that including only the creatinine sequential data in the LSTM coupled with the 115 static predictors led to a test AUC of .813 (.807-.818 95% CI). This was not an improvement over the same model with the feedforward architecture. It is possible that the 4 buckets we used were not optimal for leveraging the ability of the LSTM to learn sequential dependencies.

3.3.6 | LSTM With All Lab Values

Similar to what we found with the feedforward neural network, we found that the LSTM model with all lab values did not significantly improve upon the version without these additional lab values. This model had a test AUC of .817 (.809-.824 95% CI).

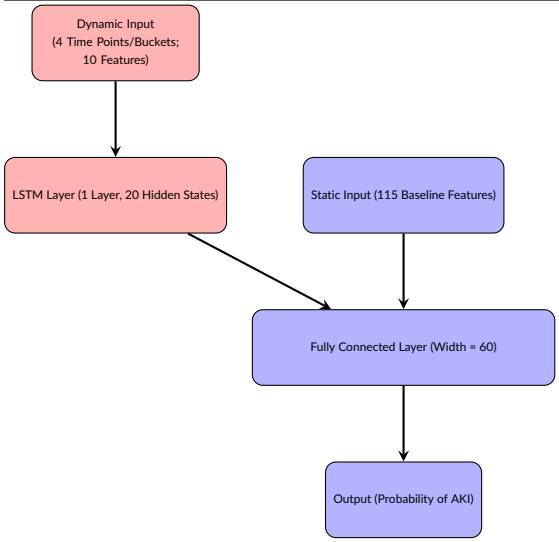

FIGURE 3 The architecture for the LSTM model.

4 | DISCUSSION

Our results are summarized in Table 2. We found that using a neural network model improves performance compared to a basic logistic regression. Additionally, incorporating longitudinal creatinine data leads to a significantly better model performance over the baseline neural network. However, adding in additional sequential lab values of substances such as blood urea nitrogen, chloride, sodium, potassium, hematocrit, hemoglobin, wbc, neutrophil absolute count, and neutrophil count, does not seem to improve model performance.

Additionally, for the features we included, our model performance did not improve using the LSTM structure compared to treating each sequential bucket of measurements as independent features.

Several further directions for our work could include measuring variable importance by removing sets of features and determining whether the model performance is significantly decreased. For example, we could remove all demographic variables or prior procedure variables to see if these significantly impact model performance.

Future work should experiment with further types of architectures, including vanilla RNNs, as well as other ways of passing in the dynamic input (such as including all sequential pre-procedure creatinine measurements instead of the average within 4 buckets). One of our difficulties was that 53% of the patients in our dataset had 3 or fewer pre-procedure creatinine measures, which means that for a majority of patients, there wasn't a long sequence of pre-procedure measurements – more data would be helpful to determine how much predictive power can be extracted from these measures. Additionally, since increasing model depth of the baseline neural network model didn't seem to improve the model results, its possible that an alternative structure such as a ResNet would bring improvements for the deeper models.

Model	# Features	Mean Testing Data AUC (95% CI)
Baseline Logistic Regression (Section 3.1)	115	.806 (.805-.812)
Baseline (Section 3.2)	115	.810 (.807-.814)
Baseline (2 Middle Layers) (Section 3.2)	115	.808 (.805-.812)
Baseline (3 Middle Layers) (Section 3.2)	115	.807 (.803-.811)
Creatinine Buckets (Section 3.3.1)	123	.819 (.815-.823)
Changing Creatinine Values (Section 3.3.2)	132	.818 (.815-.822)
All Changing Lab Values (Section 3.3.3)	285	.818 (.814-.821)
LSTM With Creatinine Buckets (Section 3.3.5)	123	.813 (.807-.818)
LSTM With All Lab Values (Section 3.3.6)	285	.817 (.809-.824)

TABLE 2 This table shows the performance of the different models.

References

[1] Nikbakhtian S, Reed AB, Obika BD, Morelli D, Cunningham AC, Aral M, et al. Accelerometer-derived sleep onset timing and cardiovascular disease incidence: a UK Biobank cohort study. European Heart Journal-Digital Health 2021;2(4):658–666.

[2] Makki N, Brennan TM, Girotra S. Acute coronary syndrome. Journal of intensive care medicine 2015;30(4):186–200.

[3] Zhang Yf, Liu Dd, Zhou Y, Lou Jz. Acute Kidney Injury in Patients with Acute Coronary Syndrome after Percutaneous Coronary Intervention: Pathophysiologies, Risk Factors, and Preventive Measures. Cardiology 2021;146(6):678–689.

[4] Huang C, Murugiah K, Mahajan S, Li SX, Dhruva SS, Haimovich JS, et al. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: A retrospective cohort study. PLoS medicine 2018;15(11):e1002703.

[5] Kuno T, Mikami T, Sahashi Y, Numasawa Y, Suzuki M, Noma S, et al. Machine learning prediction model of acute kidney injury after percutaneous coronary intervention. Scientific reports 2022;12(1):1–12.

[6] Tsai TT, Patel UD, Chang TI, Kennedy KF, Masoudi FA, Matheny ME, et al. Contemporary incidence, predictors, and outcomes of acute kidney injury in patients undergoing percutaneous coronary interventions: insights from the NCDR Cath-PCI registry. JACC: Cardiovascular Interventions 2014;7(1):1–9.

[7] Maragatham G, Devi S. LSTM model for prediction of heart failure in big data. Journal of medical systems 2019;43(5):1–13.

[8] Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:151103677 2015;.