Executive Summary: Telltale Signs of Kidney Disease: A Predictive Tool?

Amelia Spivak and Steve Manns
GitHub: https://github.com/manns79/Towards-Early-Detection-of-Kidney-Disease

Introduction

In the United States, over 130,000 people are newly diagnosed with Chronic Kidney Disease (CKD) each year. Swollen ankles or elevated blood pressure in a young person may lead to an investigation that results in a kidney diagnosis. But more often than not, kidney disease is symptom-free and only discovered accidentally through routine labs that pick up elevated levels of serum creatinine. The problem with this mode of discovery is that by the time serum creatinine levels are (confirmed to be) elevated the patient almost certainly has advanced kidney disease.

This motivates the central question underlying our project: Are there some other routine lab readings that are associated with elevated serum creatinine? If some other lab test or collection of lab tests point to a high likelihood of elevated serum creatinine, then perhaps tracking these may allow for early prediction of kidney disease before serum creatinine is elevated. Using a dataset of ICU patients with many kidney patients we seek to answer the question: Could kidney disease patients be distinguished from non-kidney disease patients in the ICU on the basis of a set of routine labs that does not include serum creatinine, the lab test whose confirmed elevated level defines kidney disease?

If we can answer this question in the affirmative in the more compromised ICU setting, then it would look very promising that our question, with the appropriate data, could be answered with machine learning methods in the affirmative in the outpatient setting.

Dataset

Our study utilized data from the fourth iteration of the Medical Information Mart for Intensive Care (MIMIC-IV) database. MIMIC-IV consists of 364,627 patients, which we categorized into two groups: kidney disease and non-kidney patients. In order to perform predictive modeling, we sought to extract the results from eight different lab tests for all of these patients. However, due to the large computational cost (over 24 hours on an HPC cluster) and the time constraints imposed by the project deadline, we restricted our analysis to a small subset of patients for which we had time to extract data. This subset consisted of 2,088 patients. Although this is a small proportion of the data that is available in MIMIC-IV, we remark that this is still significantly larger than the popular UCI machine learning repository kidney dataset, which has been featured in Kaggle competitions and scientific publications. In our data, the target is the class (1 = kidney disease and 0 = non-kidney) of the patient. All features are continuous and include albumin, bicarbonate, BUN, BUN/creatinine, calcium, chloride, creatinine, hemoglobin, and potassium. For the purpose of modeling, the value of the eight labs (excluding BUN/creatinine) was taken from the last date that a patient took all of these lab tests.

Modeling

Based on our domain knowledge and the results from our exploratory data analysis (EDA), 18 different combinations of the nine features were selected for use in the modeling process. For each of these 18 combinations, we implemented logistic regression and k-nearest neighbors classification. Of course, there are a variety of other classification algorithms one may try, but the size of our dataset precludes algorithms that are designed for large datasets. We demonstrated this point by obtaining a 99.67% accuracy and 0.66% false negative rate on the training set using a gradient boosting classifier. Such seemingly impressive metrics would make our results comparable to published research on the subject, but we do not expect for such models to generalize well to new data. For logistic regression and k-nearest neighbors classification, hyperparameter tuning was performed using a grid search. Our best model uses logistic regression with albumin, bun, and bicarbonate as features. This model was evaluated on the test set to check whether a comparable accuracy score and false negative rate was obtained. Evaluation on the test set indicated that our goal of avoiding overfitting was achieved.

Results

Our best model achieved an accuracy of 84.31% with a false negative rate of 23.38% on the test set. For the purpose of comparison, we also evaluated a logistic regression model using creatinine as the only feature on the test set. The creatinine model had an accuracy of 90.85% and a false negative rate of 14.29%. Since (confirmed) elevated levels of creatinine defines kidney disease, the latter model serves as a sort of gold standard for the purpose of comparison. With that in mind, these results are surprising. On the test data, our best model had an accuracy less than 6.5% of that of creatinine and a false positive rate about 9% greater than creatinine. We believe this result points to a promising answer for the ICU patients (patients in MIMIC-IV). Namely, albumin, bun, and bicarbonate constitute a set of routine labs that are closely associated with the creatinine level. The ICU setting is more compromised, as labs can be deeply affected by the acuteness of the situation and by the administration of IV fluids, etc. Therefore, it appears promising that, given the appropriate data, our original question may be answered in the affirmative in the outpatient setting.

Future Work

Based on the results stated above, we have identified two avenues for future work:

1. Improve the dataset:
    a. Phase 1: To confirm the surprising results we obtained for ICU patients, retrieve the lab results for more of the kidney disease patients in MIMIC-IV to perform similar machine learning tests.
    b. Phase 2: Obtain a dataset from the outpatient setting with a full set of routine labs that would allow time series analysis. Apply our candidate set of labs to this population as a first guess for a predictive tool for kidney disease. Use machine learning to augment this set with additional lab tests to achieve better accuracy.

2. Improve the modeling:
    a. Assuming a very large dataset, apply more powerful machine learning methods.