Predicting Chronic Kidney Disease

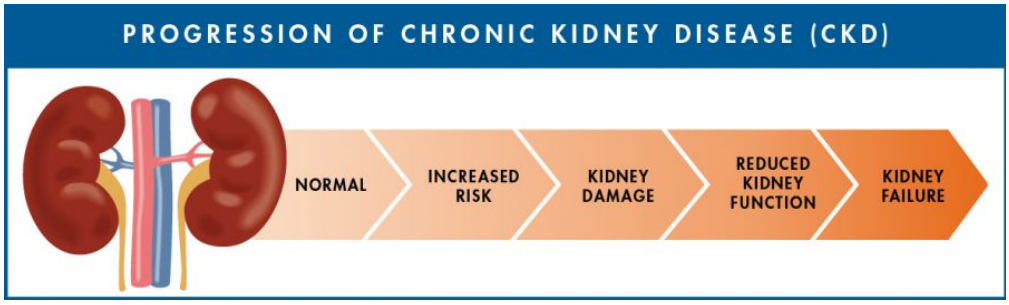
Milestone Report

Problem Statement

> *Can census data predict the hyperlocal prevalence of chronic kidney disease?*

The Problem

According to the National Kidney Foundation, 30 million Americans suffer from chronic kidney disease (CKD), a life-threatening disease that impairs the kidneys' ability to remove waste products from the body, maintain a proper balance of bodily fluids and chemicals, regulate blood pressure, produce vitamin D, and control the production of red blood cells [1]. Although CKD can be prevented and easily managed if caught in the early stages, it is often the case that it goes unchecked until the disease has progressed to more advanced stages. Recent estimates put the annual cost of treating kidney failure at $31 billion with 89,000 deaths per year [2]. Early detection of CKD can help save money, save lives, and improve the livelihood and productivity of millions of Americans.



Machine learning has been used to detect CKD[3,4] and its progression[5] using labwork data. While this is potentially great for those patients that find themselves getting bloodwork done at routine checkups or in tandem with other health-related issues, it poses two problems: first, it requires individual patients to have regular lab work done, and second, it does nothing to predict the disease at the population level.

The Centers for Disease Control (CDC) recognize chronic kidney disease as a public health concern that requires population-level surveillance and prevention[6]. However, the hyperlocal surveillance of CKD is difficult and costly. This limits our ability to effectively target public health campaigns to prevent CKD and its progression. In order to predict hyperlocal disease prevalence, readily available Census Bureau and CDC data on known and hypothesized risk factors may be used with machine learning algorithms to better identify census tracts where aggressive public health campaigns and healthcare initiatives can positively affect the early detection and treatment of CKD. This may also give insight into potential risk factors that would require further investigation.

Clients

This project was conceived to help city, county, and state public health departments decide how best to use limited resources on public health campaigns regarding the prevention and management of

chronic kidney disease. Being able to not only target high-risk areas but also gain insight into potential risk factors to target could help public health campaigns more effectively prevent chronic kidney disease and its progression. Local governments could also use the data in state and federal grants proposals in order to better make their case for the need for financial assistance.

Data

This project combines data from the Center for Disease Control's 500 Cities: Local Data for Better Health and the U.S. Census Bureau's American Community Survey 5-year Data (ACS, 2015). The 500 Cities Project contains 28,004 census tract-level observations on CKD prevalence as well as 2 unhealthy behaviors and 7 prevention measures that will be used as features.

Prevention Measure	**Unhealthy Behaviors**
Routine check-up within the past year	Sleeping less than 7 hours
Routine visits to a dentist or dental clinic	No leisure-time physical activity
Older adults aged≥65 years who are up to date on a core set of clinical preventive services (male / female)	
Cholesterol screening	
Mammography	
Papanicolaou	
Fecal occult blood test, sigmoidoscopy, or colonoscopy	

The ACS data provides 16,557 census tract-level feature variables, of which 256 were selected. The ACS variables cover an array of demographic information in the following categories:

Variable Categories	**Sub-categories**
Age	By sex
Marriage Status	Heterosexual, homosexual, children, no children
Disability	By type
Employment	Labor force participation, hours worked per week, weeks worked per year
Profession	By sector
Household Type	Family, single, by sex, 65 and over living alone
Housing Cost	Annual and as percentage of income
Insurance	Coverage by type, No coverage

Data Acquisition

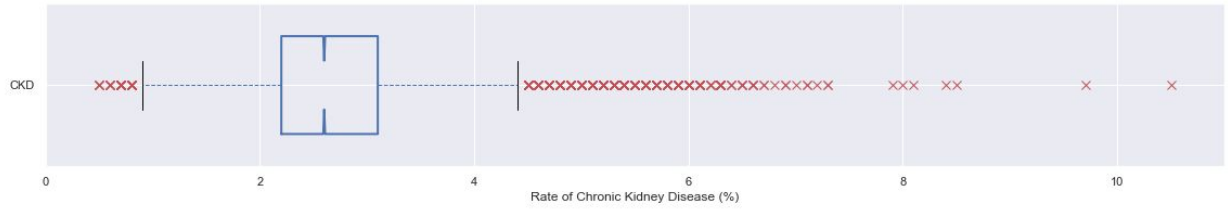
The CDC dataset was downloaded from Data.gov (https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-b32fd) on August 11, 2018.

The Census Bureau dataset was created by pulling data from the U.S. Census Bureau API. In particular, data was pulled from the 2015 American Community Survey 5-year Data (https://www.census.gov/data/developers/data-sets/acs-5year.2015.html). Since this project required tract-level data from all 50 states, it required 50 separate API calls. The API requires the desired variables to be explicitly called, so a dataframe was first created from the variable JSON (https://api.census.gov/data/2015/acs/acs5/subject/variables.json). The unwanted variables were systematically dropped from the dataframe. The variable ID column was then converted to a list and used in the API call. All 50 datasets were then merged into one dataframe, on which all further data wrangling was performed.

Data Wrangling

The CDC data was paired down to include only tract-level data. It was then pivoted so that each row represents a census tract and each column a variable. All chronic disease variables other than prevention services and chronic kidney disease were dropped from the dataset. This was due to the fact that I wanted the final model to be easily reproducible for more than just 2015. Since the CDC dataset has only ever been released once, it seems unlikely that this will provide reliable predictive power in future years. Preventative services were kept in order to test the potential hypothesis for how to reduce rates of chronic kidney disease.

The data set was then merged with the population count, state, city, and geolocation data from the CDC data. Since the ACS dataset uses Tract FIPS numbers as unique identifiers, these numbers were extracted from the CDC's UniqueID column and saved as TractID (the same variable name given in the ACS dataset). Most of the data were converted to numeric data types, with state and city converted to categorical features. Outliers were kept since they seemed to correlate with actual data. There were only a few variable missing data and at most they were missing 1.97%, but any missing data were imputed using the mean value of the relative variable.

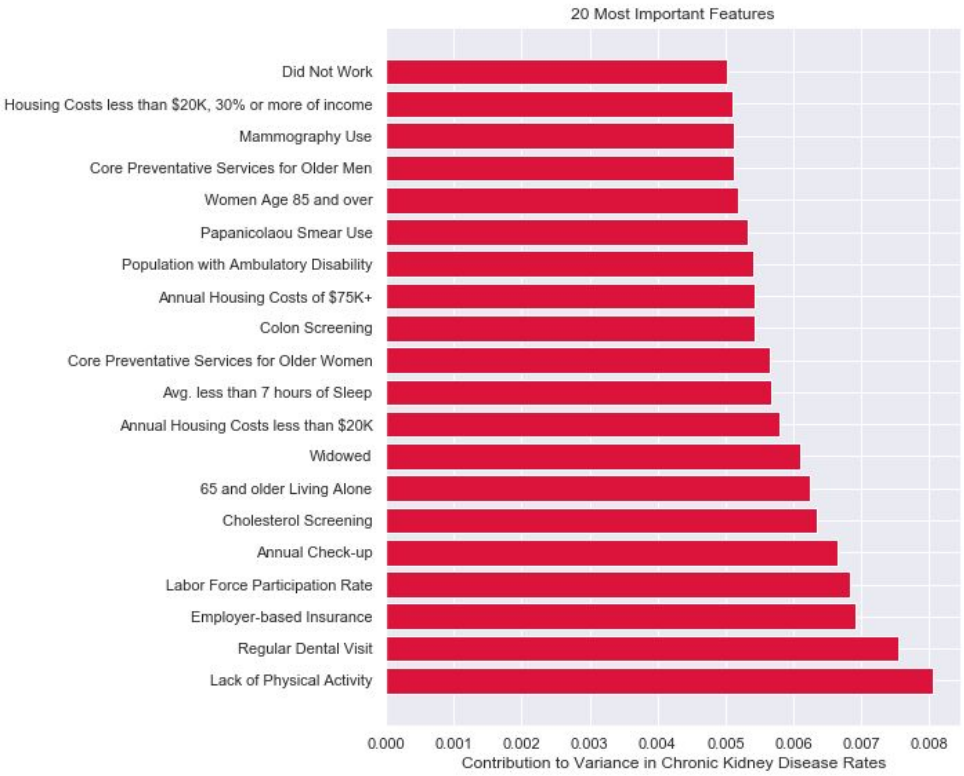


The ACS data was first converted to numerical data types. Two variables pertaining to the sexuality of unmarried couples were found to be completely empty and therefore dropped. Missing values, which were marked with negative values, were converted to NaN. Columns missing more than 80% of their data were dropped. This brought the number of variable down to 237, and the remaining missing values were imputed with the variable mean. The column names were then changed to be more descriptive (and then truncated to conserve space).

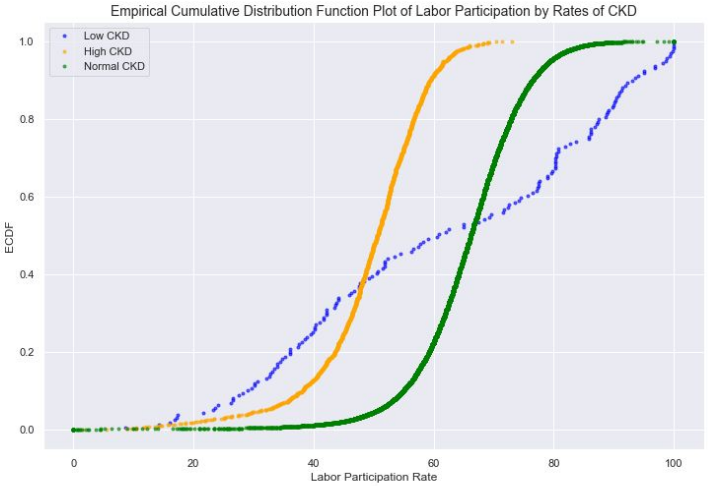
The CDC and ACS dataframes were merged and duplicate rows were dropped. This brought the final shape of the data to 27,408 observations and 252 variables.

Exploratory Data Analysis

For the purpose of exploratory data analysis and statistical inference, I first wanted to focus on finding a few different types of features using an extra trees classifier to identify some of the most important predictors. This classifier fit 75 random decision trees (with a max depth of 20) on various sub-samples of the data in order to predict the features' contributions to the variance in rates of CKD. If you run this classifier with the other chronic diseases from the CDC dataset included, the top 10 features are always other chronic health diseases (with the top 3 always being stroke, diabetes, and coronary heart disease). Since this data is not regularly available at the census tract level, I decided to eliminate these features. However, it is worth exploring the epidemiology of CKD as it relates to other chronic diseases.



Without the chronic disease data, we see a more diverse set of features with unhealthy behaviors and preventative measures topping the list. While this data is also not regularly available, it may provide some useful information into how to lower future rates of CKD and we will build models without such data in order to provide a more useful prediction method. Two factors that we will examine further are the type of insurance and the labor force participation



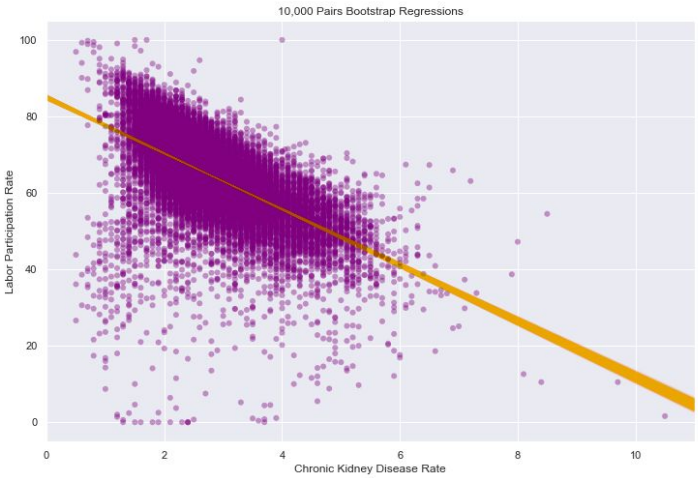
rate (both of which appear in the top 4 features). Another pattern of note from the feature importance ranking is the number of female-based variables that appear. This may be due to the fact that women (14.94%) suffer from chronic kidney disease at higher rates than men (12.35%)[7].

Whether the Labor Force Participation Rate had a significant impact on rates of CKD was examined by splitting the tracts into three groups of high, low, and normal rates of CKD (high and low were defined as being 2 standard deviations away from the mean). From the ECDF we can see that there are not a lot of census tracts with low rates of CKD (only 159 of 27,408) and the low rates do not appear to be normally distributed (in fact, they seem to be bimodal). However, the distribution does appear to be quite separate and descriptive statistics back this up, although there are large degrees of standard error.

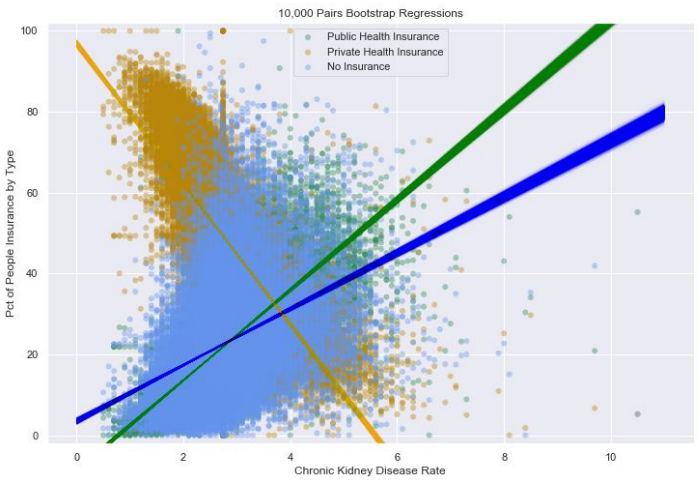
	Mean CKD Rate	Mean Labor Participation Rate	Standard Error of Labor Participation
Low Rates of CKD	0.939%	61.62%	25.18%
High Rates of CKD	4.917%	49.32%	9.65%

Assuming a null hypothesis that there is no difference in mean labor participation rates between groups with low and high rates of CKD and a significance level of 0.01, the 99% confidence interval for the difference of means is between -5.13 and +5.25.

. With our sample difference of means of 12.30, we calculate a z-score of 6.11, which translates to an estimated p-value of essentially 0.0. In addition, our 99% confidence interval for the difference of means calculated from the sample data is between 7.084 and 17.705, which does not include zero. For these reasons, we reject the null hypothesis and find that there may be a difference in mean labor participation rates between groups of high and low CKD.



A bootstrap analysis only further solidifies this rejection of the null hypothesis by providing a 99% confidence interval of the difference of mean assuming the null hypothesis to be between -0.051 and +0.052, with a bootstrap sample mean difference in means of 12.30 and a z-score of 615.79, giving a p-value of virtually 0.0. This implies that low rates of labor participation can be correlated with higher rates of CKD.

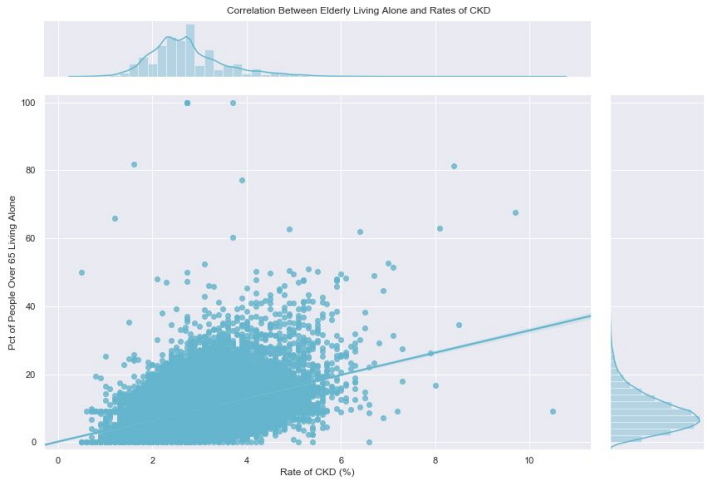


Next, we examine whether the type of insurance coverage (public, private, or no insurance) may have an effect on rates of CKD. We will assume a null hypothesis that the type of insurance does not affect the rate of CKD. Examining a scatter plot of the data with 10,000 bootstrapped slopes, we

can see that private health insurance is very strongly negatively correlated with rates of CKD and nearly orthogonal to the slope of public health insurance. Interestingly, public health insurance appears to be more positively correlated with CKD than no health insurance. This is corroborated by a bootstrap analysis of the difference of means and difference of means slopes.

	Sample Difference	99% Confidence Interval	z-score	p-value
Difference of Mean Slope	4.23	(4.228,4.236)	2751..27	0.0
Difference of Pearson r	0.2814	(0.281, 0.282)	4285.0	0.0

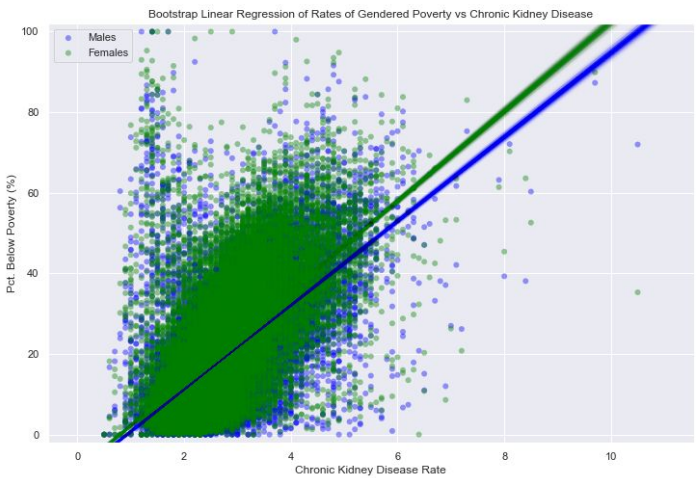
The difference between the slope of public and no insurance is likely due to the fact that older people are more likely to have public insurance (Medicare) than no insurance as compared to younger populations, and since age is strongly correlated with rates of CKD, there is a potential collinear effect. In addition, people with stage 5 kidney disease are eligible for Medicaid. A bootstrap analysis of private health insurance shows a 99% confidence interval for the slope between -17.78 and -17.13, indicating a strong negative correlation. We, therefore, reject the null hypothesis and find evidence to believe that the type of insurance may have an effect on rates of CKD. While it makes intuitive sense that the lack of health insurance would lead to higher rates of CKD, it seems implausible that an increased average age would alone account for the positive correlation with CKD.



The third variable examined is the percentage of people over the age of 65 who are living alone. Examining the scatterplot with the regression line, we can see there appears to be a positive correlation. The hypothesis was first tested using the sample data and then 10,000 bootstrapped samples. The results are shown in the table below. Since the data appears to be skewed right for both variables, the Spearman ρ was also calculated. Spearman ρ at 0.4033 is slightly lower than the Pearson r, but with a p-value of 0.0, it is still well below the significance level. Since our mean slope and correlation coefficient both remain above 0 in the 99% confidence interval, we will reject the null hypothesis.

Sample Data	
Slope	3.29
Pearson r	0.4459
p-value	0.000
Standard error	0.0405

Bootstrap Samples		Confidence Interval
Mean Slope	3.29	(3.137, 3.454)
Pearson r	0.4459	(0.428, 0.463)
r^2	0.1989	(0.183, 0.215)



In order to examine potential interaction effects between economic factors and sex, we can examine the differences in the rate of CKD between males and females in relation to their respective rate of poverty. To test the null hypothesis that gender does not play a role in how poverty affects the rate of CKD, the difference in mean slopes and mean Pearson r coefficient using bootstrap analysis were computed. The results lead us to reject the null hypothesis and assume that gender combined with poverty status may provide more insight into the rate of CKD than poverty alone.

	99% Confidence Interval	Mean Difference	z-score	p-value
Difference of Mean Slope	(0.6604, 0.6695)	0.6650	386.58	0.0000
Difference of Mean Pearson r	(0.0096 0.0101)	0.0098	110.38	0.0000

In conclusion, we can see that many of the variables, such as labor force participation rate, types of insurance, gender, poverty status, and seniors living alone, may help to explain some of the variance in rates of chronic kidney disease. By combining these features in a regression model, we should be able to predict the local rate of CKD.

References

1. National Kidney Foundation. About Chronic Kidney Disease. National Kidney Foundation A to Z Health Guide. https://www.kidney.org/atoz/content/about-chronic-kidney-disease. Last accessed October 30, 2018
2. National Kidney Foundation. End Stage Renal Disease in the United States. National Kidney Foundation website. https://www.kidney.org/news/newsroom/factsheets/End-Stage-Renal-Disease-in-the-US. Last accessed October 30, 2018.
3. Misir, R., Mitra, M., & Samanta, R. K. (2017). A Reduced Set of Features for Chronic Kidney Disease Prediction. *Journal of pathology informatics*, *8*, 24. doi:10.4103/jpi.jpi_88_16
4. Soltanpour Gharibdousti, Maryam & Azimi, Kamran & Hathikal, Saraswathi & H Won, Dae. (2017). Prediction of Chronic Kidney Disease Using Data Mining Techniques.
5. Tangri N, Stevens LA, Griffith J, et al. A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. *JAMA.* 2011;305(15):1553–1559. doi:10.1001/jama.2011.451
6. Centers for Disease Control and Prevention. About the CKD Initiative. Chronic Kidney Disease Initiative. https://www.cdc.gov/kidneydisease/about-the-ckd-initiative.html. Last accessed November 1, 2018.
7. Centers for Disease Control and Prevention. Age-adjusted prevalence of CKD Stages 1-4 by Gender 1999-2012. Chronic Kidney Disease (CKD) Surveillance Project website. https://nccd.cdc.gov . Last accessed November 6, 2018.