From Interviews to Blood Tests: Using Statistical and Machine Learning Models to Predict Kidney Function

Priyanka Nanayakkara, Ryan Masson, Joyce Ling
University of California—Los Angeles, Northwestern University, University of California—Berkeley

Introduction

•The kidneys serve the vital function of blood filtration. Chronic Kidney Disease (CKD) is when the kidneys cannot perform adequate filtration, leaving a patient's body with waste build-up.[1]

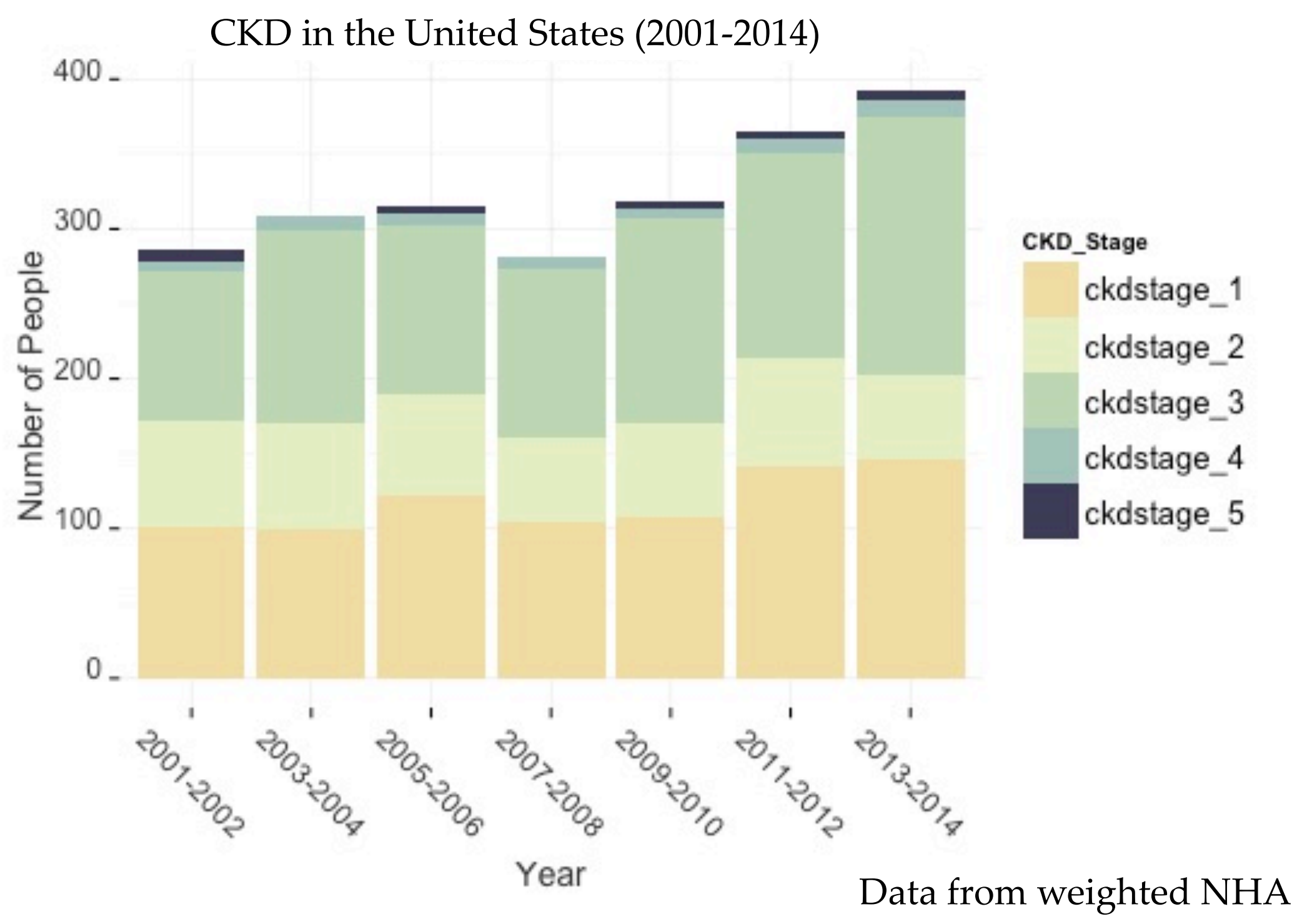
•CKD affects approximately 26 million adults in the United States. This does not include those who are at risk for the disease.[2]

•In order to measure kidney function, we looked at glomerular filtration rate (GFR), "the best estimate of kidney function."[2]

•The National Health and Nutrition Examination Survey (NHANES) dataset contains over 78,000 records, with over 100 variables (including both interview questions and medical examination results).

Our goals:
•Firstly, we wanted to see if simple interview questions could predict kidney function. The motivation behind this type of prediction was to find a low-cost way of flagging those at risk of CKD without patients necessarily having to have medical tests, or blood drawn.
•Secondly, we wanted to use both statistical modeling and machine learning methods to make accurate predictions for GFRs, and see how these two methods can complement one another.

Visualizing CKD Trends



Data from weighted NHANES subsample

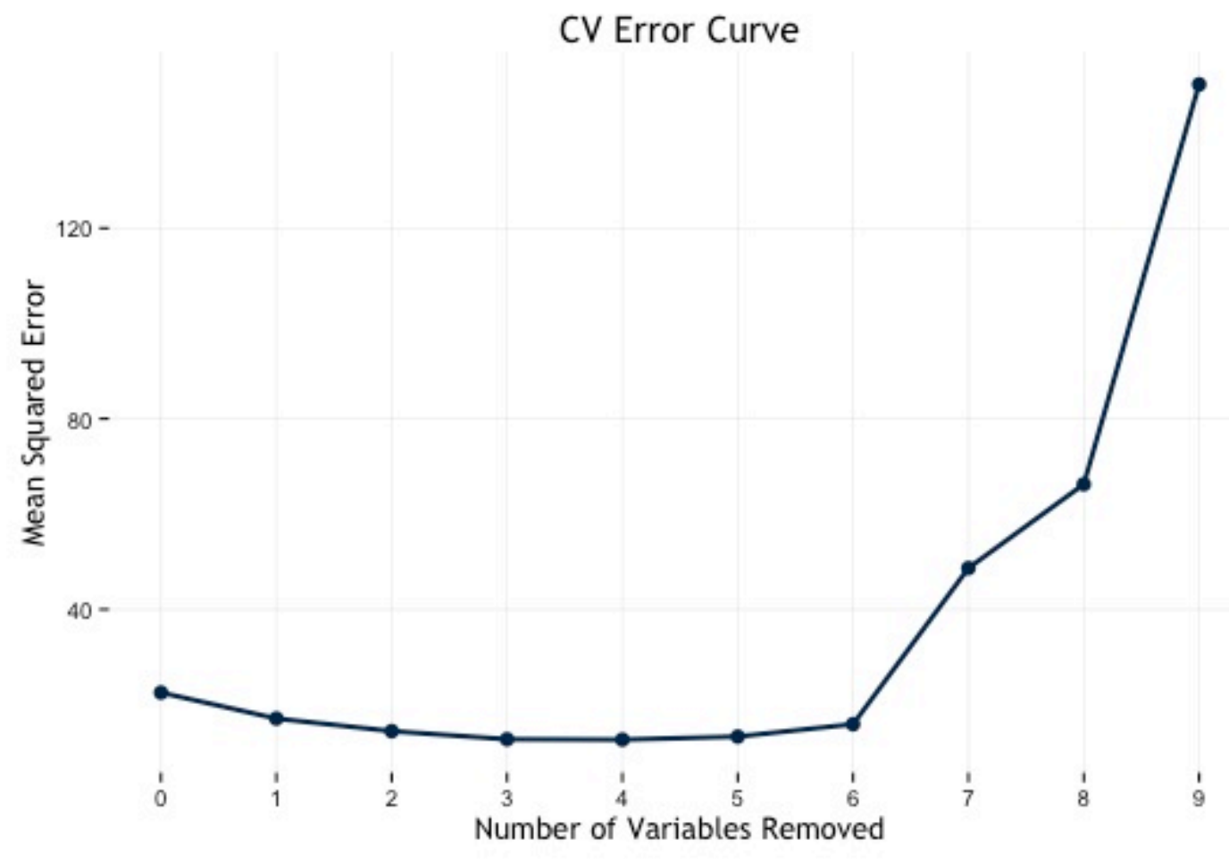
Methods

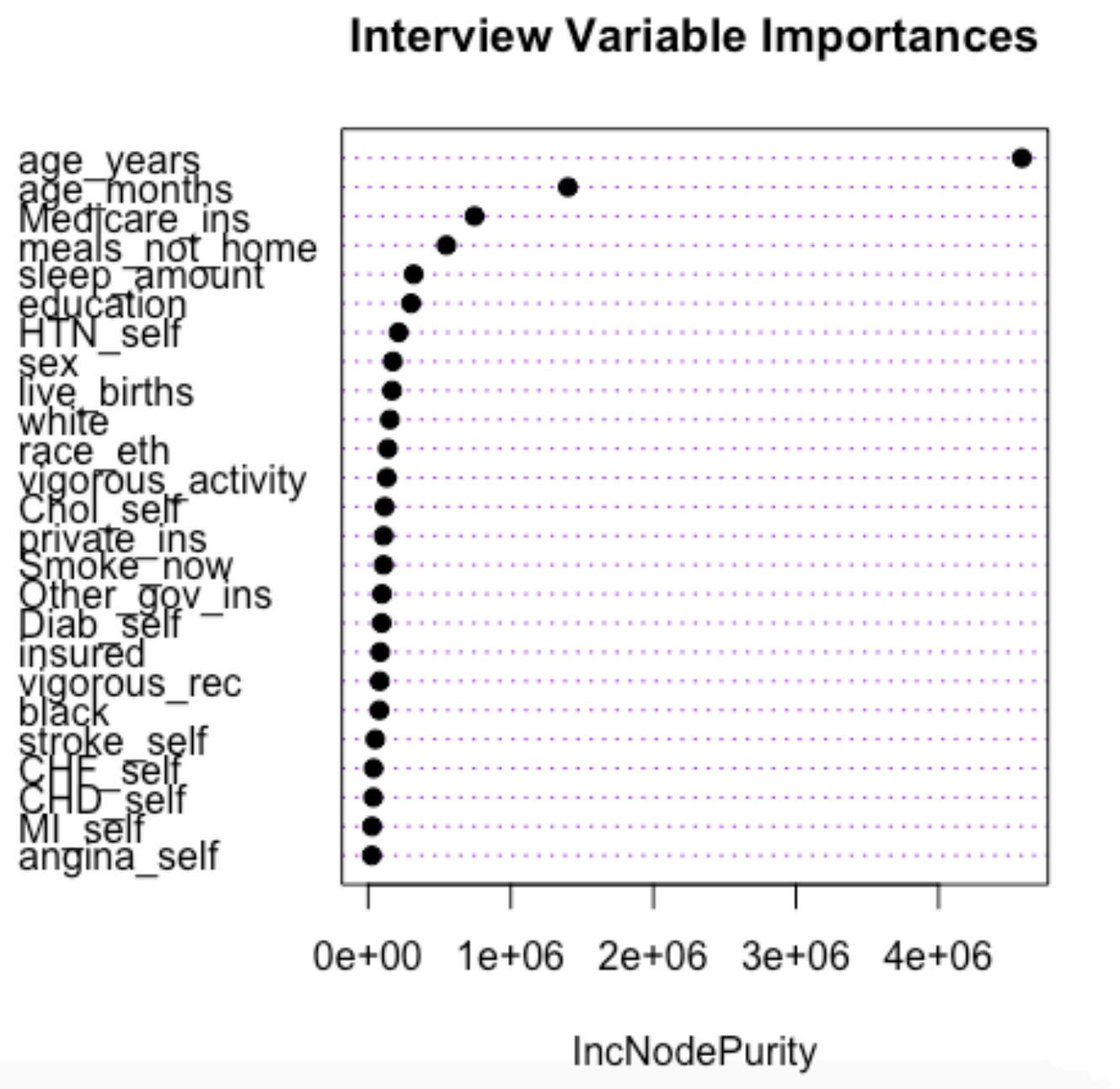
Bootstrapping to account for weighting
While the NHANES dataset is not a simple random sample, we can still use it to extrapolate to the entire United States population using the given weights for each record. We used the weights to find the proportion of the population that each record represents, and used this proportion to sample 20,000 records with replacement. We continued to use this method of sampling throughout the project whenever we needed a sample from NHANES.

Methods

Linear Regression Model
To find a set of variables that predict GFRs in a linear regression model, we first started out with an initial set of ten variables. We selected these variables based partially on a few different iterations of importance rankings (obtained through random forest method below) and whether they could be easily asked during an in-person or paper survey.

We then found the optimal variables out of the initial set using MSE as the criteria for backwards selection . To find MSE values at each step of backwards selection, we implemented 10-fold cross validation. The MSEs initially decreased, and hit a minimum when four variables were removed, demonstrated by the error curve below. Hence, we chose to discard four variables and were left with six.





Random Forest Model
Using the same top ten variables as covariates, we fitted a random forest model using the R package "randomForest." We implemented 10-fold cross validation to assess model performance. Each tree in the forest split at each level on a randomly selected sample of three variables, to account for the possible correlation of variables (mtry = 3). The total number of trees used in the forest was 400 (ntree = 400), meaning that each of the 400 random samples of the training data was fit to an individual regression tree. The overall prediction of the forest was the average prediction of the regression trees (this process is called "bootstrap aggregation" or "bagging," and lowers the high variance of an individual regression tree to give a better prediction).

Results

Linear Regression Model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

eGFR ~ Meals Outside of Home + Sleep Amount + Smoking + Race (as White or Non-White) + Hypertension (self-reported) + Age

The linear regression model to the right was fitted to the 20,000 observation sample.

```
Coefficients:
                 Estimate  Pr(>|t|)
(Intercept)     141.13328   < 2e-16 ***
meals_not_home   -0.03388   0.56623
sleep_amount     -0.38747   0.01936 *
Smoke_now        -0.48454   0.05929 .
white            -5.18984   < 2e-16 ***
HTN_self          1.51644   0.00104 **
age_years        -0.88238   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
Multiple R-squared:  0.5169
```

Results

MSE & RMSE

	Linear Regression	Random Forest
MSE	991.68	203.71
RMSE	23.70	14.25

We obtained the above MSE values by performing 10-fold cross validation on ten random samples of the NHANES dataset in the same manner we obtained the dataset referred to in the Methods section.

Discussion

Model Based Recommendations
•Linear Regression Model: We would like to perform an in-depth residual analysis to see if any of the predictors do not have a linear relationship with the response variable (GFR). Additionally, we would like to look further into why there are relationships between the predictors and CKD, and especially focus on examining the repeatedly significant predictors.

•Random Forest Model: The main improvement is reaching the optimal tree depth while running the algorithm over all variables to get their importance rankings. Because of our computing power restraints, we could not go deep enough. Using a parallelized random forest package or a more powerful computing cluster could help us with this next step.

Sampling and Accounting for Survey Weights
Additionally, when we used our bootstrapped sample with n = 20,000 to build a linear regression model, we had much smaller MSE values (at least by a factor of 10) than when we used a sample with n = 40,000. We suspect this might be because with a 40,000 sample, more observations from the original dataset were repeated. In any case, this major discrepancy in MSE values requires further investigation.

Randomness
Randomness played a surprising large role in our linear regression model. When we tested the linear regression model on the ten 5,000 record samples, selected by different random seeds, the MSE values ranged from 36.35 to 4855.64. We suspect that this points to the lack of robustness of the model, but we would like to further understand why the model does not maintain its predictive power with different samples.

References

1. Staff, Mayo Clinic. "Chronic Kidney Disease." *Mayo Clinic*. Mayo Foundation for Medical Education and Research, n.d. Web. 13 July 2016.
2. "About Chronic Kidney Disease." *National Kidney Foundation*. National Kidney Foundation, Inc., n.d. Web. 13 July 2016.

R and R Packages
A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
B. Hamner (2012). Metrics: Evaluation metrics for machine learning. R package version 0.1.1.
H. Wickham.(2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.
M. Kuhn et al. (2016) caret: Classification and Regression Training. R package version 6.0-70.
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Acknowledgements

We would like to thank Dr. Yanming Li, Dr. Jian Kang, and Dr. Kevin He for mentoring us and answering all our questions.
We would also like to thank those involved in organizing the Big Data Summer Institute; we've had a wonderful summer!
The NHANES dataset came from the Centers for Disease Control and Prevention (CDC).