Final Report

Organ Transplant Patient Wait Time Analysis

Dae Hyun Kim

Disclaimer

"The data reported here have been supplied by the United Network for Organ Sharing (UNOS) as the contractor for the Organ Procurement and Transplantation Network (OPTN). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government."

This research was performed based on OPTN data as of July 15, 2021.

Introduction

Organ transplant question comes up frequently in medical school interviews to access candidates' reasoning skill. The fundamental problem associated with the organ transplant is the high demand and low supply. Due to the low availability of adequate organs, decisions need to be made on who should get the organ. From patients' perspective, this may be a life or death decision not to mention the financial burden of medical bills as they aimlessly wait for an organ.

One interesting fact about organ transplant system in the US is that the distance between donor and recipient matters. Currently, the US is divided into 11 regions and the donated organ is first distributed within the region. This practice is probably implemented to ensure the quality of organ during the long-distance travel. However, as the technology develops with better transporting system, this regional boundary may seem unfair for some. In particular, there is a state discrepancy in which patients in certain states will more likely obtain an organ.

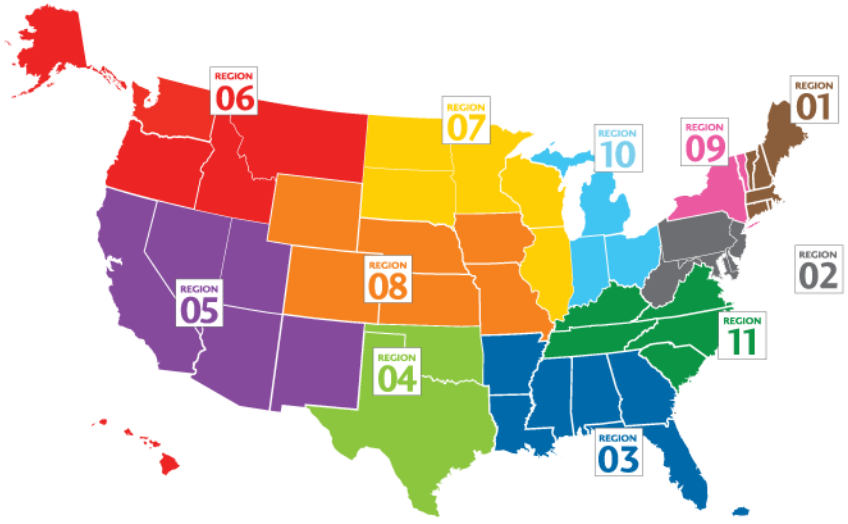

Figure 1. the 11 US organ transplant regions from UNOS.

I analyzed organ transplant data from OPTN to verify if indeed there are regional disparities. Another main objective for this analysis is to derive a predictive model to estimate the patient wait time based on the recipient's profile. Since health data cannot include the diversity of each case and personality, the prediction model was not very successful. However, based on RandomForestRegression model, I was able to identify the calculated patient's survival rate, patient's age of admission, and the distance from donor hospital from treatment center as the three most important variable in wait time prediction.

Data Wrangling

The raw data from OPTN contains information on all waiting list registrations and transplants of organs since October 1, 1987. The delimited text file was 13GB large so it was not possible to read the data to begin with. For this analysis I selected kidney transplant only because that was the most frequent transplant in the United States. I started reading 1000 rows of the data to understand the format and data entry. After practicing data wrangling from the sample data, I read the whole kidney data by chunks and selected the date on and after year 2011. The result was 400,542 rows of data. I also inspected the descriptions of 490 columns and selected 59 as relevant.

Exploratory Data Analysis

This part of the process overlapped with the data wrangling step because I didn't do a good job cleaning the data previously. I only focused on reducing the data to a manageable size and didn't consider the data quality. As I was inspecting the histogram of each column, I was also cleaning the data simultaneously.

There were two notable observations while exploring the data. First of all, the number of wait listing registrations steadily increased except the year 2020. This may have to do with the COVID-19 pandemic. Also, note unusually low number for year 2021 is due to the fact that record was obtained July 15 of 2021. By roughly multiplying by 2, the estimated number of waitlist registrations should be compatible to the year 2019.

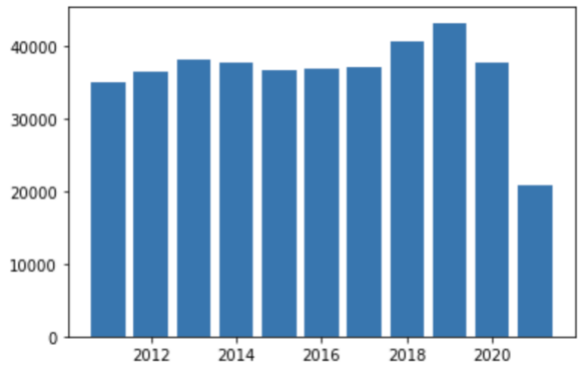

Figure 2. the number of waitlist registrations by year.

Second observation was the difference among states. Below are two bar graphs of number of waitlist registrations (blue bars) and the number of kidney transplants (orange bars). CA denotes California and OH means the state of Ohio.

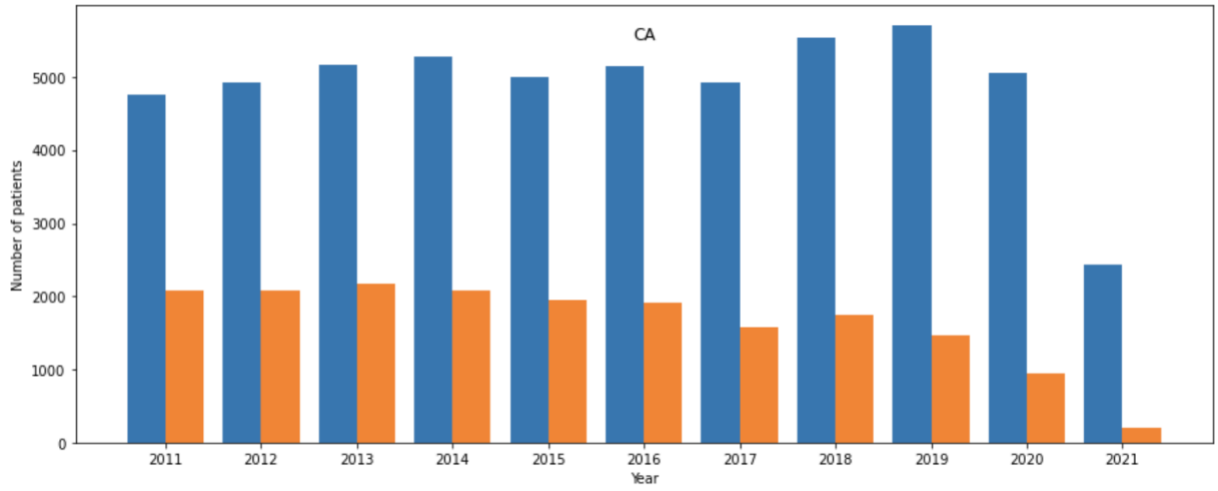

Figure 3. registrations vs transplants in California (0.338).

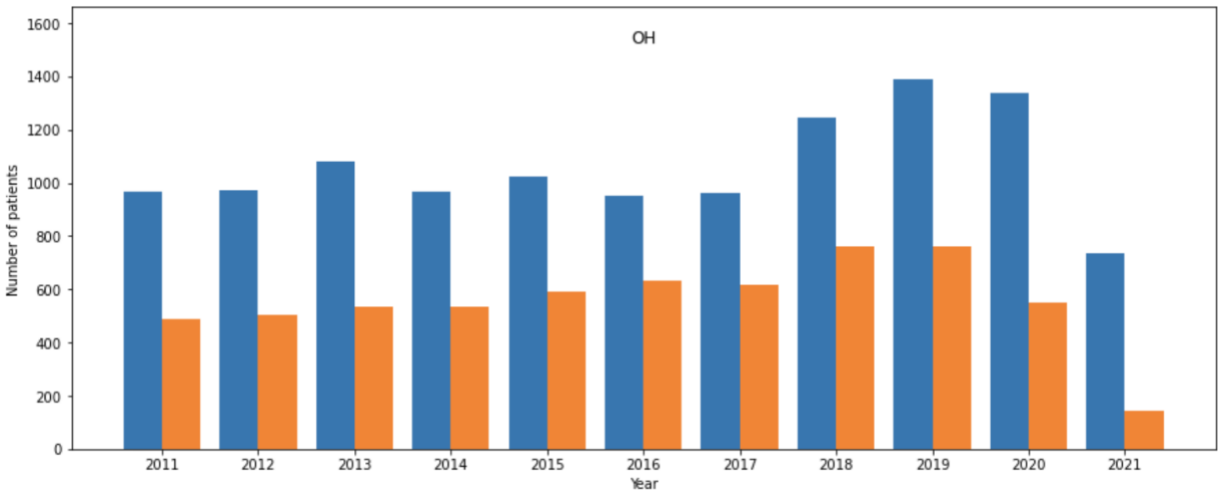

Figure 4. registrations vs transplants in Ohio (0.525).

Visually, there is a higher chance for people in Ohio to gain access to kidney compared to California. Some may argue that these graphs represent the regional discrepancies in the US organ transplant system. Numerically, the highest ratio was the state of Nebraska with 0.630 while the lowest ratio was the state of Delaware with 0.287. However, one thing to note is the difference in number of patients. California has patients over 5000 per year whereas Ohio is less than the third of that number. Others may disagree because already California takes the most of available organs.

Finally, the heatmap for selected features were mapped. Unfortunately, for the wait days variable, no significant correlation was observed. This is probably due to the fact that my data is not linear.

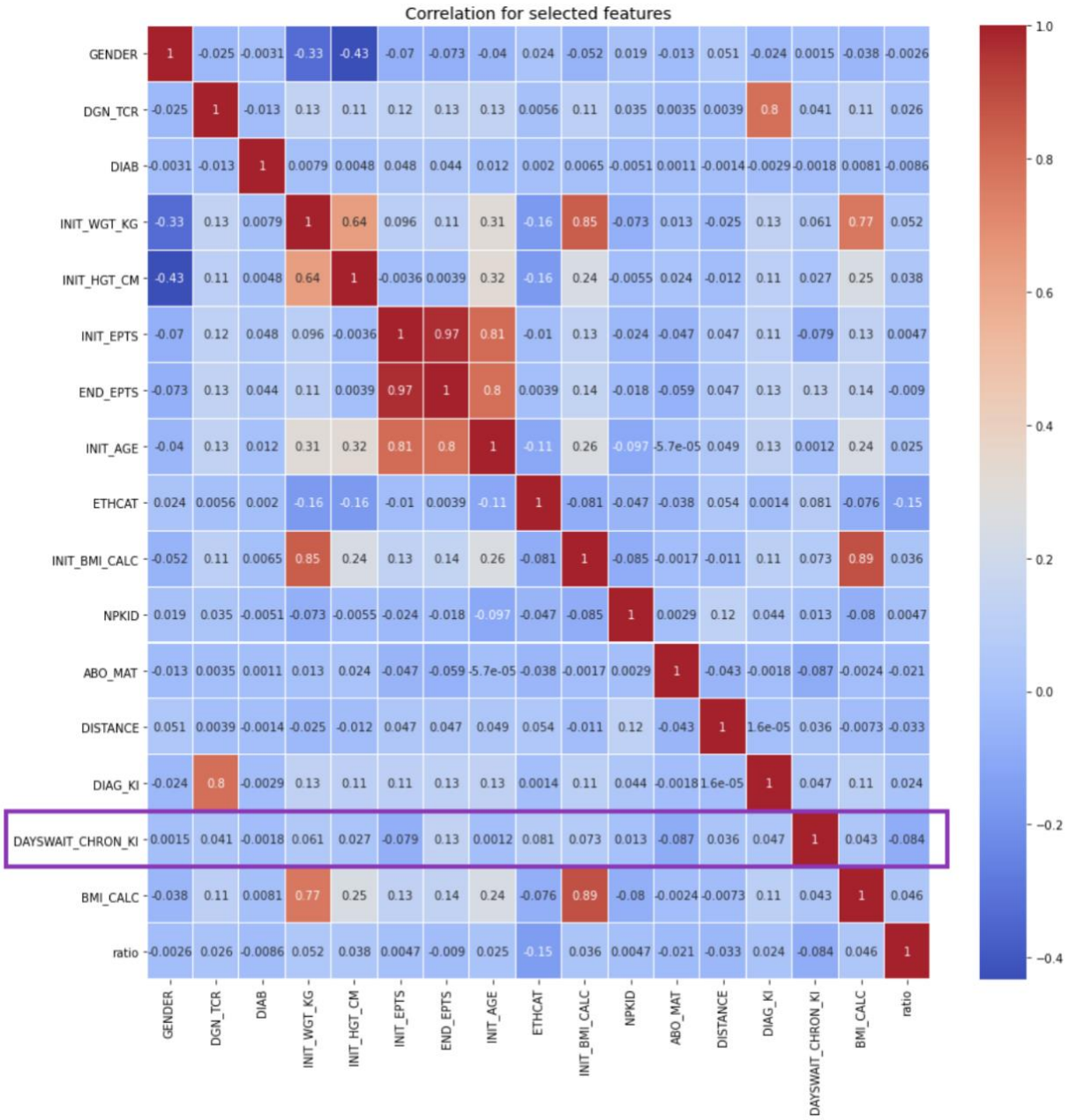

Figure 5. heatmap for selected features. Highlighted

Preprocessing

Before using models, I further cleaned the data. Especially, I reduced the rows with only transplant patients because many patients are still waiting to for an organ transplant. Also, I excluded the donor information from the features as much as possible because for a patient registering, they won't know about the donor at all. The resulting data was 166,404 rows and 16 columns. From here, I replaced missing values, dropped columns with too many missing values, and created dummy variables for categorized features. For numeric values, I scaled

using standard scaler. The target variable for this analysis is the wait time for each patient based on the profile they have at the time of registration and some other variables (such as distance).

Model Selection

I ran several regression techniques to create a model that can predict the wait time of a patient. I started with dummy mean predictor that only predicts out of the mean of the training set. This will be used as a baseline comparison for every other model. Then I looked into linear regression. I tuned the hyperparameter such as the limiting k (number of features), and using elastic net model with ridge and lasso hyperparameters. The results were recorded in Table 1. As suspected from the heatmap, no linearity was present, and the accuracy was terrible.

As an alternative, I looked into RandomForest regression model. By altering the tree size and max depth, I was able to increase R^2 value from 0.09 to 0.29. The main issue with the RandomForest was the runtime of the random grid search. Each search took more than 2 hours to perform, so I stopped after 5 trials and reported the best result.

Algorithm	Training Set			Test Set		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Dummy mean	0	612.0	473.5	0	611.4	471.8
Linear Regression (k)	0.115	575.7	441.1	0.116	574.8	439.1
Linear Regression (Elastic Net)	0.138	568.3	437.5	0.137	567.8	435.9
RandomForestRegression	0.287	517.0	388.3	0.265	524.1	393.3

Table 1. model metrics.

The above table shows that RandomForestRegression is the best algorithm, and there is a room for improvement as I perform more hyperparameter searches.

Takeaways

I was happy that I increased the R^2 up to 0.287. I read online that for many health data, there are too many unexpected variables because we are dealing with human data. Therefore, the R^2 value of 0.3 may not be such a bad thing. Unfortunately, this is not really the case for my analysis because my target variable is the wait time for waitlisted patients and the average error for my prediction is above one year. With a more powerful computer, I may have tuned the hyperparameters and improved my predictive model, but I don't think continually searching is not very efficient.

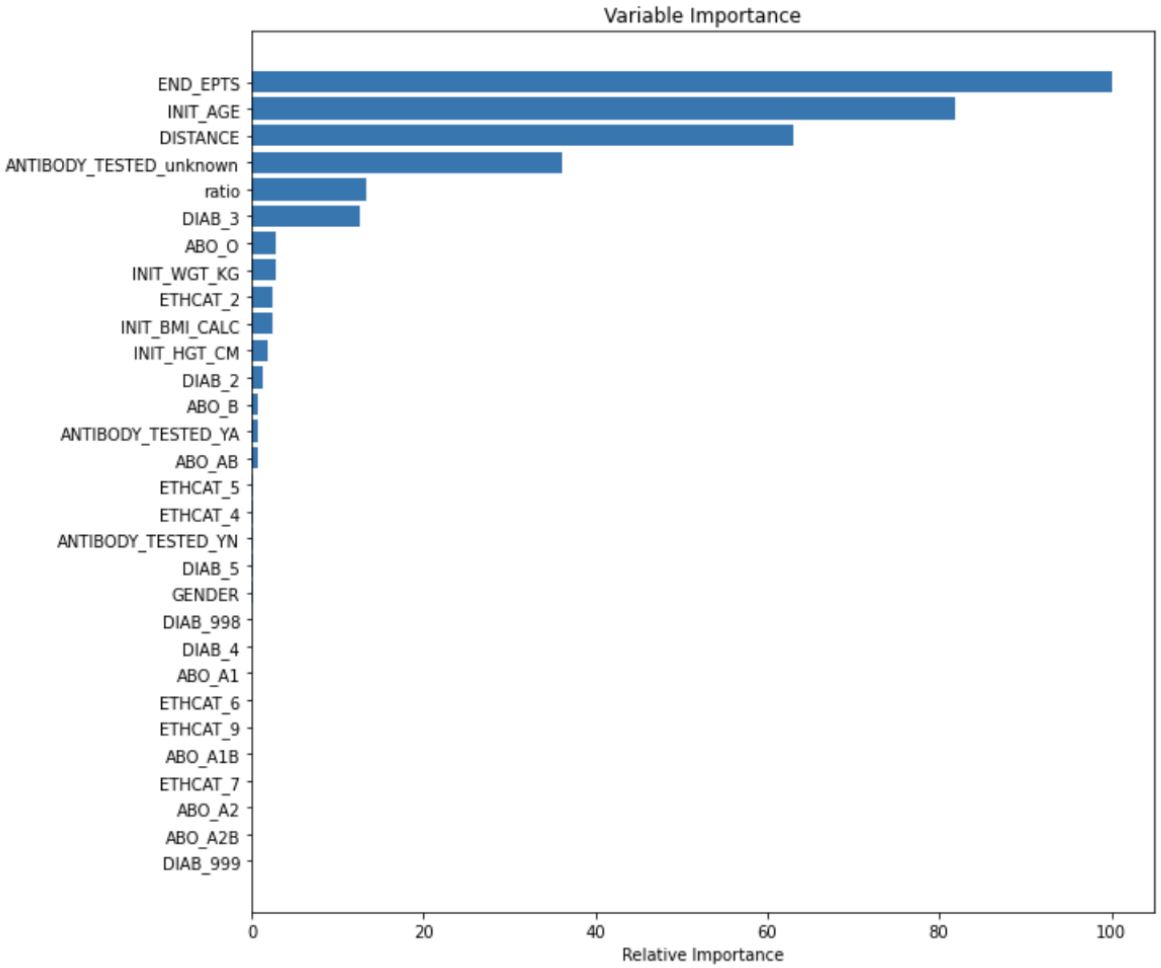

Figure 6. feature importance in RandomForestRegression.

On the other hand, while observing the feature importance, I found that three most relatively important features were END_EPTS, INIT_AGE, and DISTANCE. To clarify END_EPTS is the calculated survival rate for each patient. INIT_AGE is the initial age at the time of registration, and DISTANCE is the distance from donor center to the transplant procedure center. This was interesting because it means that the distance was more important than some of these features which include initial diagnosis and diabetes status. This partially support my initial suspicion about distance being a deciding feature for organ transplant which may seem unfair for some patients. Nonetheless, the low accuracy of this model should not be overlooked.

Future Direction

The main issue with my analysis is the low predictive power and consumability. In short, my model is useless. There are several ways to improve my model.

1. Continue tune hyperparameters – seems inefficient and time consuming
2. Analyze data by year – keeping year column may reveal time sensitive information.

Also, my mentor suggested several other methods to improve my analysis.

3. Cutting data certain years from now – result in better model for recent data.
4. Change the problem to classification – setting a time frame such as within a week, within a month, within a year, and more than 2 years maybe more beneficial for patients as well as easier for machine to predict.

I will most likely continue with option 2 to see if the important feature changes by year. I believe option 4 will be the best to increase my accuracy and create a consumable product, but I will have to redo the modeling part from the scratch. I will probably work on this when I have more time.