# Predicting the Improvement of NBA players

Fatemeh

AlvaniMay 1,

2019

## 1. Introduction

### 1.1 Background

Founded in 2014, San Francisco-based Bayes Impact is a group of experienced data scientists assisting nonprofits in tackling some of the world's heaviest data challenges. Since it's founding, Bayes has helped the U.S. Department of Health make better matches between organ donors and those who need transplants, worked with the Michael J. Fox Foundation to develop better data science methods for Parkinson's research, and created methods to help detect fraud in microfinance. Bayes is also developing a model to help the City of San Francisco harness data science to optimize essential services like emergency response rates. Through organizations like Bayes, data science has the power to make a significant social impact in our data-driven world.

### 1.2 Problem

Making better matches between organ donors with a lot of data is not an easy job, this project aims to predict whether and how Donor owners can match the needs as per their Age, Gender Blood group and all possible other criteria.

### 1.3 Interest

Obviously, Medical teams would be very interested in accurate prediction of best organ matches, for competitive advantage and health values. Doctors, Donors and Patients may also be interested.

## 2. Data acquisition and cleaning

### 2.1 Data sources

Most Patients and Doors stats are the majority of data, the historical log of previously successful operations as well helps to predict the best donors matches.

## 3 Data cleaning

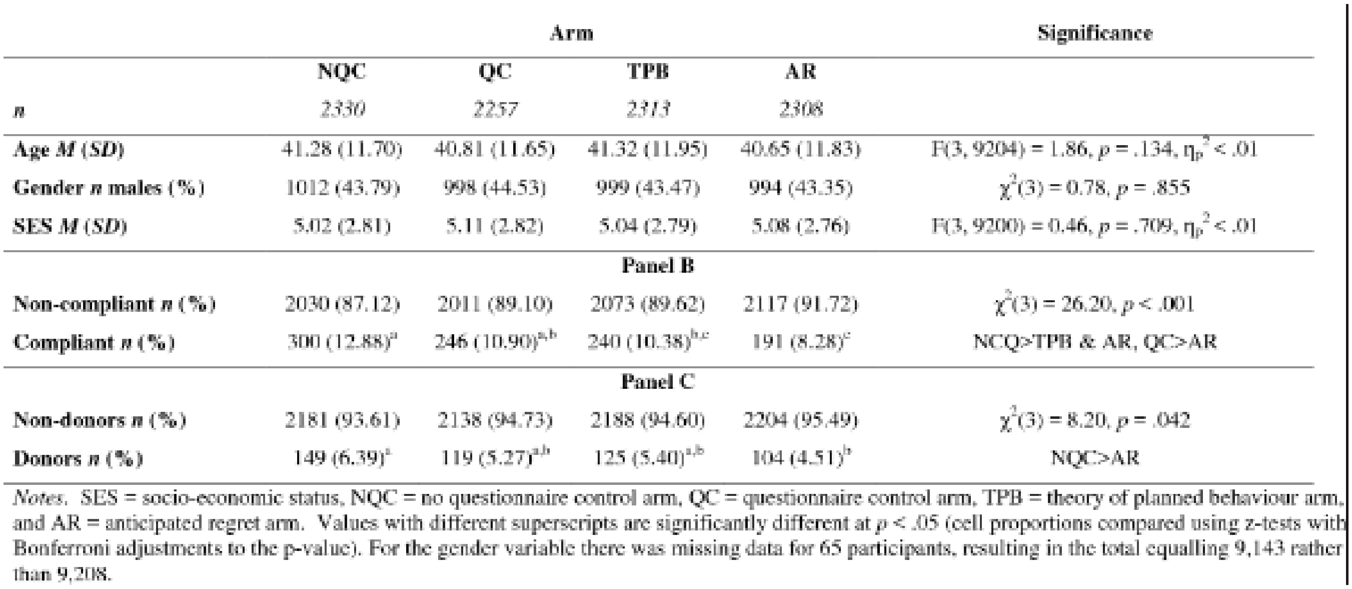
Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values from earlier operations, because of lack of record keeping. I decided to only use data from 1999 and after, because of  fewer missing values and Operations was a lot different in the early years from today's Procedure.

There are several problems with the datasets. First, patients were identified by their names. However, there were different with the same names, which cause their data to mix with each other's. Though it was possible to separate some of them based on the years, organ types, and organ they Donated,  Therefore, patients with duplicate names were removed.

Second, multiple entries existed for patients who has unsuccessful donates. This cause their Medical data to represent multiple samples with incomplete data. I wrote script to extract total stats for these patients, and discarded partial rows.

## 4 Feature selection

After data cleaning, there were 54,725 samples and 32 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. For example, there was a feature of the number of patients collected, and another feature of the number of organs donated. These two features contained very similar information, with the difference being that the former feature increased with Operation time, while the latter feature did not. Such total vs. rate relationship also existed between other features. These features are problematic for two reasons: (1) A patient's certain donations were duplicated in two features. (2) operation time were duplicated in multiple features. In order to fix this, I decided to keep all features that were rates in nature, and drop their cumulative counterparts (Table 1).



| | Arm | | | | Significance |
|---|---|---|---|---|---|
| | **NQC** | **QC** | **TPB** | **AR** | |
| *n* | *2330* | *2257* | *2313* | *2308* | |
| **Age *M* (*SD*)** | 41.28 (11.70) | 40.81 (11.65) | 41.32 (11.95) | 40.65 (11.83) | F(3, 9204) = 1.86, $p = .134$, $\eta_p^2 < .01$ |
| **Gender *n* males (%)** | 1012 (43.79) | 998 (44.53) | 999 (43.47) | 994 (43.35) | $\chi^2(3) = 0.78$, $p = .855$ |
| **SES *M* (*SD*)** | 5.02 (2.81) | 5.11 (2.82) | 5.04 (2.79) | 5.08 (2.76) | F(3, 9200) = 0.46, $p = .709$, $\eta_p^2 < .01$ |
| | | | **Panel B** | | |
| **Non-compliant *n* (%)** | 2030 (87.12) | 2011 (89.10) | 2073 (89.62) | 2117 (91.72) | $\chi^2(3) = 26.20$, $p < .001$ |
| **Compliant *n* (%)** | 300 (12.88)$^{a}$ | 246 (10.90)$^{a,b}$ | 240 (10.38)$^{b,c}$ | 191 (8.28)$^{c}$ | NCQ>TPB & AR, QC>AR |
| | | | **Panel C** | | |
| **Non-donors *n* (%)** | 2181 (93.61) | 2138 (94.73) | 2188 (94.60) | 2204 (95.49) | $\chi^2(3) = 8.20$, $p = .042$ |
| **Donors *n* (%)** | 149 (6.39)$^{a}$ | 119 (5.27)$^{a,b}$ | 125 (5.40)$^{a,b}$ | 104 (4.51)$^{b}$ | NQC>AR |

*Notes.* SES = socio-economic status, NQC = no questionnaire control arm, QC = questionnaire control arm, TPB = theory of planned behaviour arm, and AR = anticipated regret arm. Values with different superscripts are significantly different at $p < .05$ (cell proportions compared using z-tests with Bonferroni adjustments to the p-value). For the gender variable there was missing data for 65 participants, resulting in the total equalling 9,143 rather than 9,208.

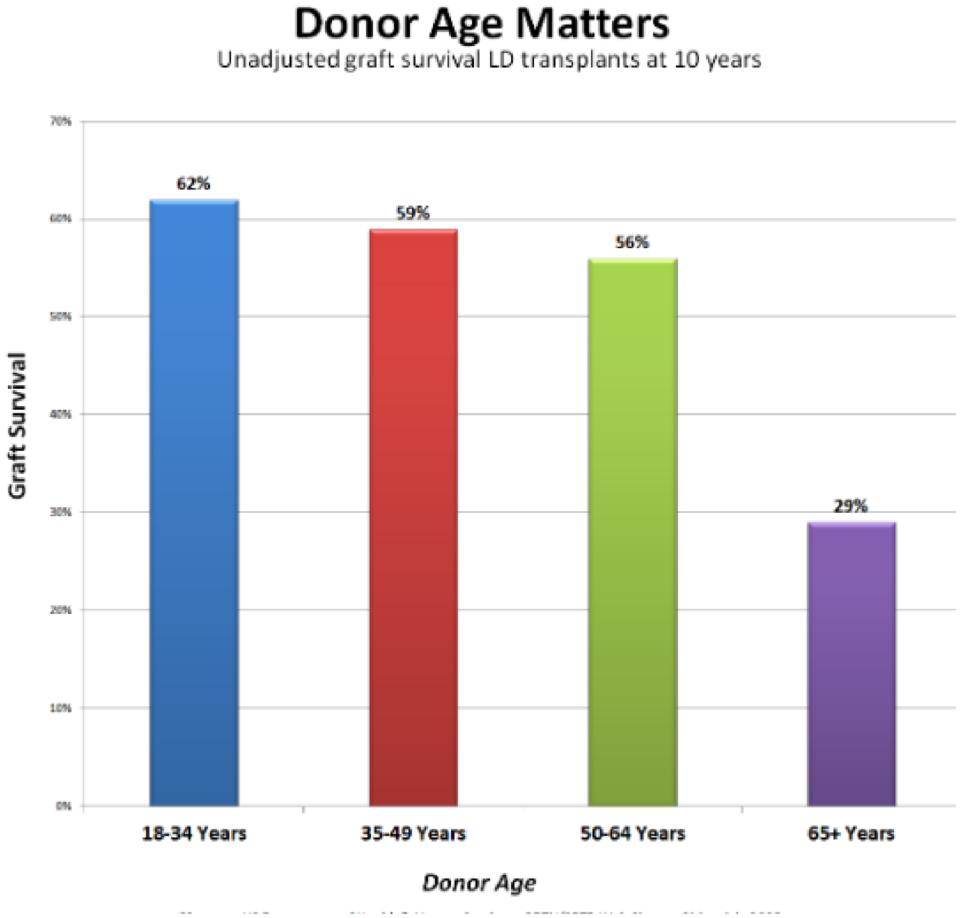
## 5 Exploratory Data Analysis

### 2.2 Calculation of the best donor variable

Operations improvement year over year was not a feature in the dataset, and had to be calculated. I chose to calculate the difference of successful operations between two consecutive years as the target

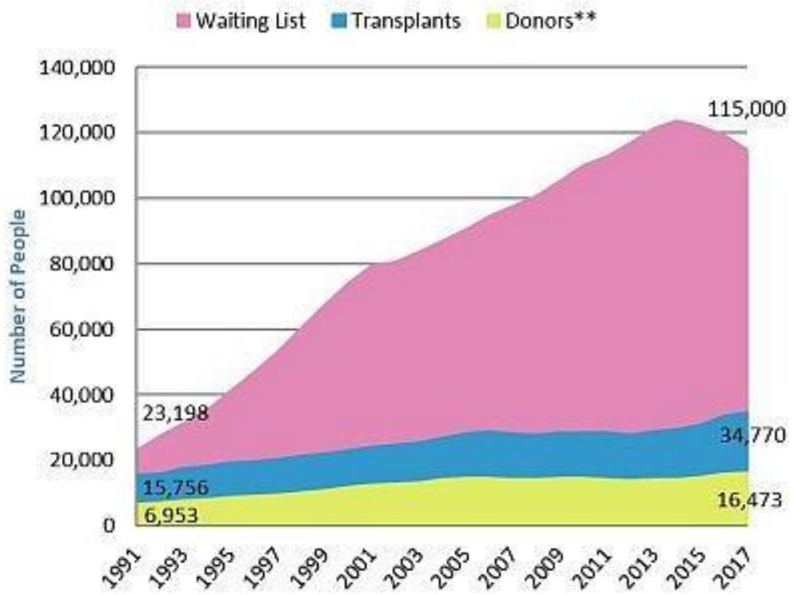
### 2.3 Relationship between improvement and age

It is widely accepted that younger donors are more likely to have successful operations, and it was indeed supported by our data. median improvement declined as patients age increased



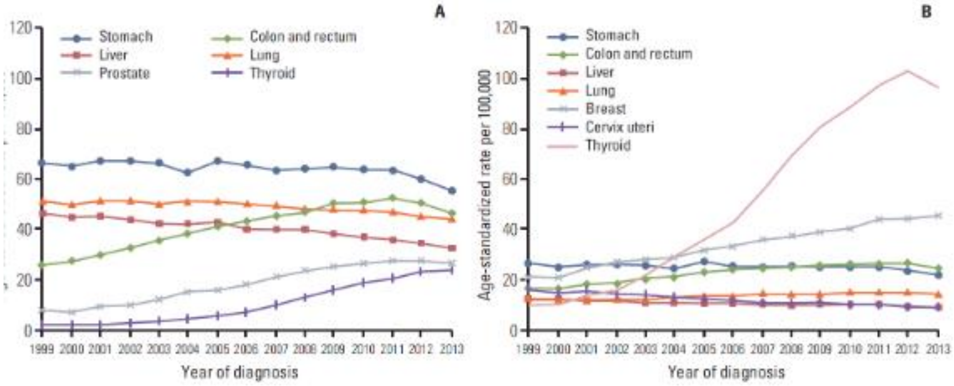
### 6 Relationship between improvement and Operations

I observed a negative relationship between player improvement and the games played (Figure 5). If a good player missed significant numbers of games, it was probably because of injury, which might have negatively impacted his performance. He might return to his former form next season, and therefore improve. Players who played fewer than 50 games were more likely to improve than those who played more than 50 games. (z-test, p<0.001, difference of mean=1.3).



## 7 Solution to the problems

The reason behind these problems were the uneven distribution of Donors improvement, in that patients with little improvement/decline were more common than patients with big improvement/decline. Therefore, the models tried to prioritize minimizing errors on patients with little improvement/decline when RMSE was used as the evaluation metric. My solution to this problem was to assign weights to samples based on the inverse of the abundances of target values. Using this method, all models predicted target values with similar range and distribution as the actual target values



## 8 Conclusions

In this study, I analyzed the relationship between Patients successful operations and their Body data. I identified age, blood type, gender, organ types that affect the operation. I built both regression models and classification models to predict whether and how much donation will be successful. These models can be very useful in helping Medical doctors in a number of ways. For example, it could help identify best donor matches, operation success prediction best matches etc.