# Method

## Dataset overview

The dataset used in this study comes from a Parkinson disease study conducted by six U.S. medical centers: Georgia Institute of Technology, National Institutes of Health, Oregon Health and Science University, Rush University Medical Center, Southern Illinois University, and University of California Los Angeles. All institute made use of the *at-home testing device* (AHTD) to record patient behaviors remotely.

A total of 52 patients were put on a six-month tracking period where the patients were asked to complete automatically prompted vocal protocols. These protocols consist of six phonation assignments, requiring the patient to pronounce the constant vowel sound “ahhhhh” at their comfortable pitch for four times and two times with twice the initial loudness.

Based on the collected data, multiple feature extracting algorithms were applied. The following table was listed out all relevant criteria and their definition:

|  |  |  |
| --- | --- | --- |
| **MDVP Jitter(%)** | **MDVP Jitter(Abs)** | **MDVP: RAP** |
| MDVP Jitter as a percentage | MDVP Jitter in microseconds | MDVP Relative Amplitude Perturbation |
|  |  |  |
| **MDVP: PPQ** | **Jitter: DDP** | **MDVP: Shimmer** |
| Period Perturbation Quotient | Difference of Differences between cycles, divided by the average period | Local shimmer |
|  |  |  |
| **MDVP: Shimmer(dB)** | **Shimmer: APQ3** | **Shimmer: APQ5** |
| Local shimmer in decibels | Three point Amplitude Perturbation Quetient | Five point Amplitude Perturbation Quetient |
|  |  |  |
| **Shimmer DDA** | **NHR** | **HNR** |
| Difference between consecutive difference between the amplitudes of consecutive periods | Noise-to-Harmonics Ratio | Harmonics to Noise Ratio |
|  |  |  |
| **RPDE** | **DFA** | **PPE** |
| Recurrence Period Density Entropy | Detrended Fluctuation Analysis | Pitch Period Entropy |

The term MDVP stands for the often-used KayPentax multidimensional voice program.

## Analysis Method

This study consists of four main parts: 1) EDA of the data, 2) A Full Regression Model of the dataset, 3) ANOVA blocked by different phonation task, 4) LSD analysis blocked by patient gender.

1. EDA

During the Exploratory Data Analysis, we double checked the missing values of each feature, and drew up a correlation matrix of all the features used presented in the dataset.

1. Full Regression model

Following the EDA, a full regression model of the dataset was performed. Based on the results gotten from the regression model, we were able to extract the significant features using the mallows’ Cp criterion. Then, we performed an ANOVA test on the full regression model and the reduced regression model with only the extracted significant features.

1. ANOVA blocked by different phonation task

As introduced in the dataset overview, the data acquiring consists of 6 tasks. Presented with such characteristic, we chose to block this factor. By assigning each phonation task with their corresponding number, we were able to 1) run regression on all six methods, 2) extract significant features of each method, 3) run ANOVA for each method of the full and the reduced regression model.

1. LSD Analysis blocked by patient gender

Lastly, within each analysis of the different phonation task, we also performed a LSD analysis on the two patient genders. By blocking patient’s gender, we hope to investigate the difference between male patient Parkinson disease and female patient Parkinson disease.