#SASGF

VIRTUAL

SAS® GLOBAL FORUM 2021

Prediction of Parkinson’s Disease with a Regression-based Model Using Vertical Ground Reaction Force Data

Jin Ge, Columbia University

Jin Ge is a graduate student majoring in Biostatistics at Columbia University. Her research interests include panel data, high dimensional data.

She has broad experience in applying SAS to clean and analyze different kinds of healthcare data, such as self-reported depressive symptom data, COVID-19 confirmed case rate data and socioeconomic data. Her research experience uses statistical learning methods to explore the potential association hidden in the data and demonstrates the interpretable result to people with different backgrounds.

Jia Guo

Columbia University

Jia Guo is a graduate student majoring in Biostatistics at Columbia University.

His research interests include machine learning, deep learning and statistical genetics.

Tianchen Xu

Columbia University

Tianchen Xu is a graduate student majoring in Biostatistics at Columbia University.

His research interests include electronical health data, microbiome data, etc.

Mengyu Zhang

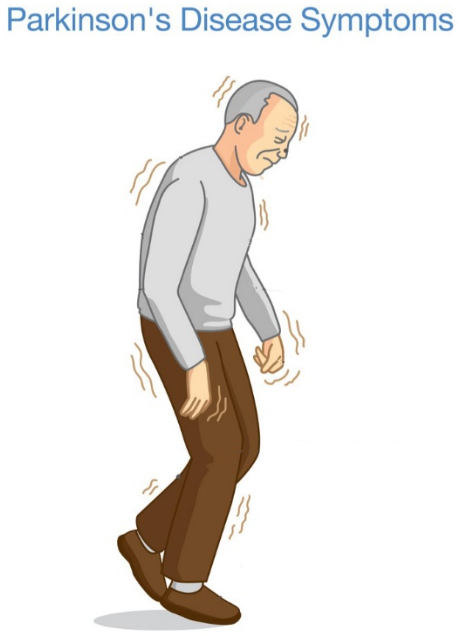
Columbia University

Mengyu Zhang is a graduate student majoring in Biostatistics at Columbia University and going to pursue a PhD degree in Biostatistics.

She is crazy about statistical learning and feature engineering..

Parkinson’s Disease

What is PD



- long-term degenerative disorder of the central nervous system
- A very common disease
- Movement or motor related difficulties such as tremor, bradykinesia, rigidity, and postural instability

Parkinson's Disease

Diagnosis

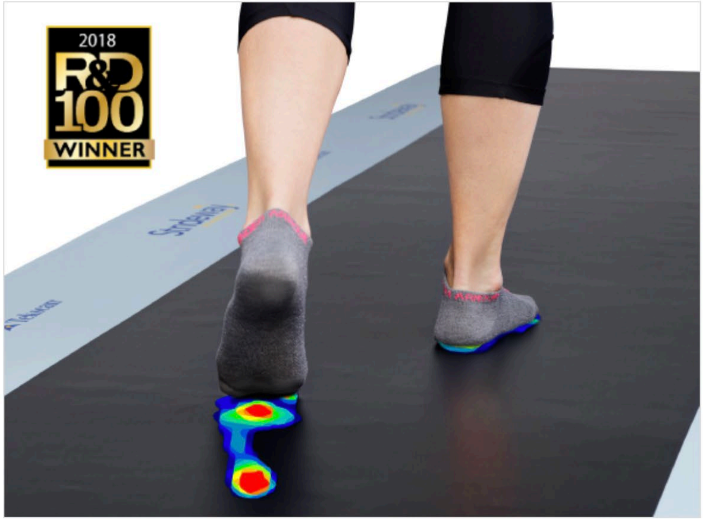
- Making an accurate diagnosis of Parkinson's disease can be complicated.
- No conclusive screening or test, patients with very early Parkinson's disease may not meet the clinical diagnosis criteria
- Gold standard: subjective clinical evaluation
  - U.K.'s Parkinson's Disease Society Brain Bank
  - International Parkinson and Movement Disorder Society
- Quantitative gait analysis system

Parkinson's Disease

Gait Analysis System

- wearables and non-wearables

Tekscan offers two in-shoe solutions depending upon your data collection needs.



Strideway is our platform based gait analysis solution.

- Ground Reaction Force (GRF)

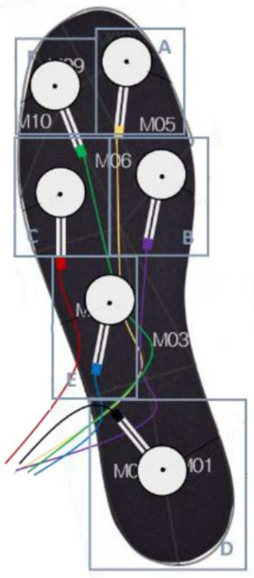
Parkinson’s Disease

Data

- Research Resource for Complex Physiologic Signals

PhysioBank PhysioNet PhysioToolkit

  - 92 patients with idiopathic PD and 73 healthy controls
  - 8 sensors each foot; walked 2 minutes on level ground
  - 100 signals per second per sensor
  - sum of the 8 sensor outputs
  - 18 time series measurements in total
- Demographic information, measures of disease severity



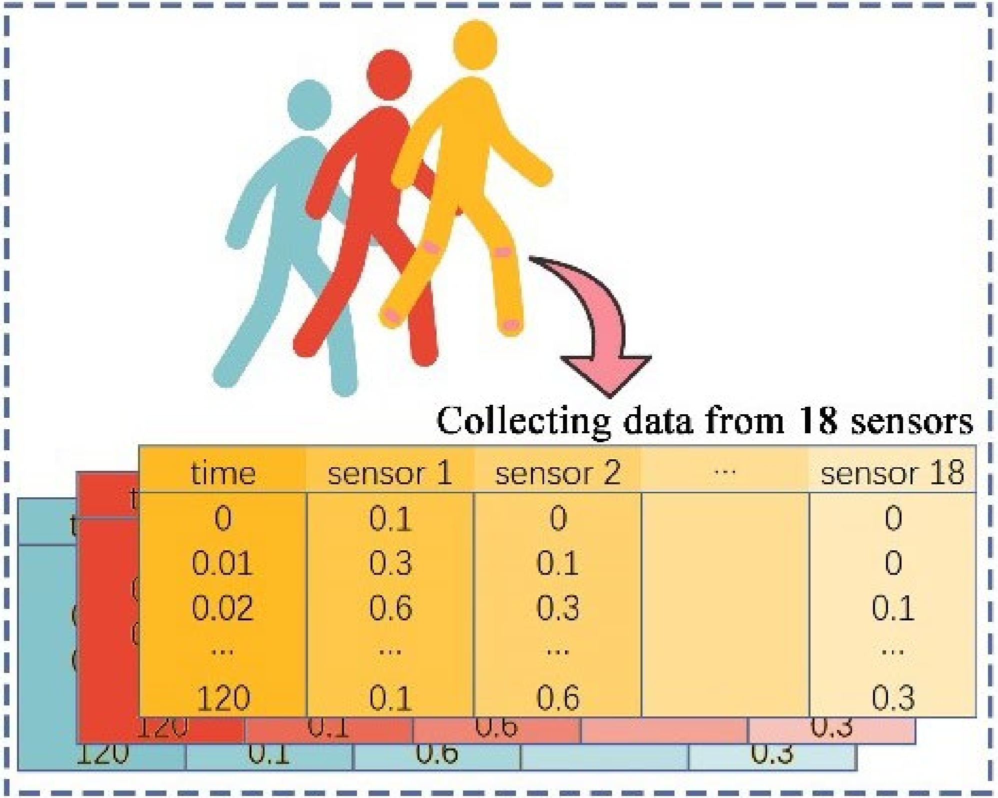
Parkinson's Disease

Research Question

What is the most influential features, including demographics and motor related biomedical signals, for PD prediction?

Method

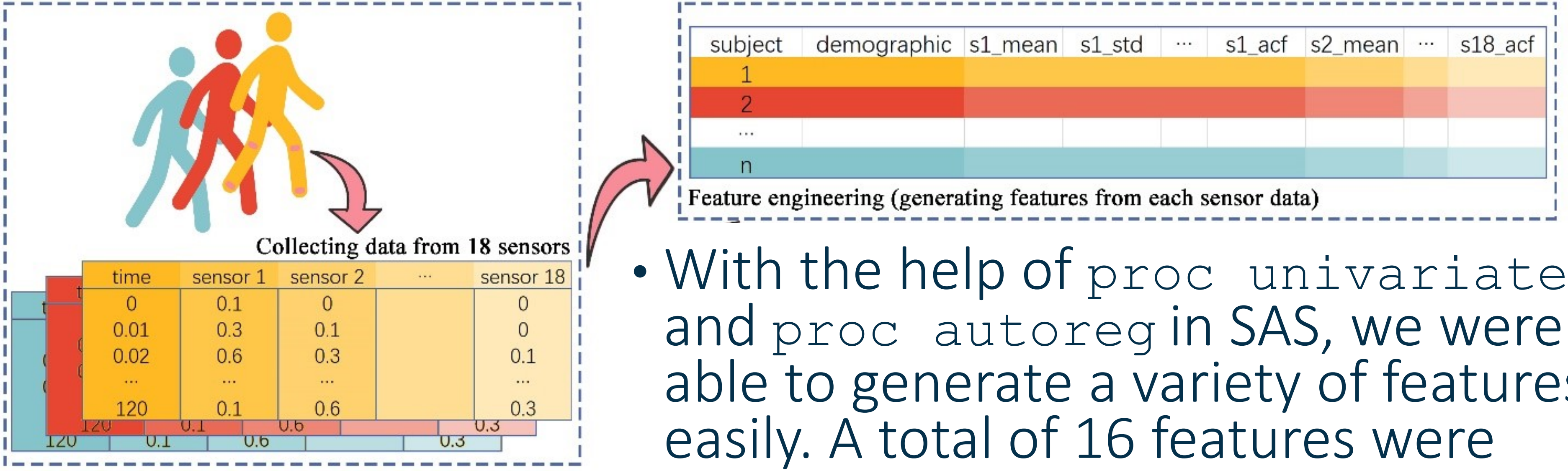
Data Collection



- We collected time series data from 18 Vertical ground reaction force (VGRF) sensors.
- We need to extract relevant digital biomarkers from these raw sensor data for further downstream analysis.

Method

Feature Extraction & Data Engineering



- With the help of `proc univariate` and `proc autoreg` in SAS, we were able to generate a variety of features easily. A total of 16 features were calculated [Snyder et al., 2020].

Method

Feature Extraction & Data Engineering

- The following features were calculated for each of the 18 time series variables [Snyder et al., 2020]:

1) Location measures: mean value, mode, median, first/third quartile, 95th quantile;
2) Dispersion measures: standard deviation, interquartile range, coefficient of variation;
3) Shape measures: skewness, kurtosis;
4) Autocorrelation with lags 10, 30, 60, 100, 120.

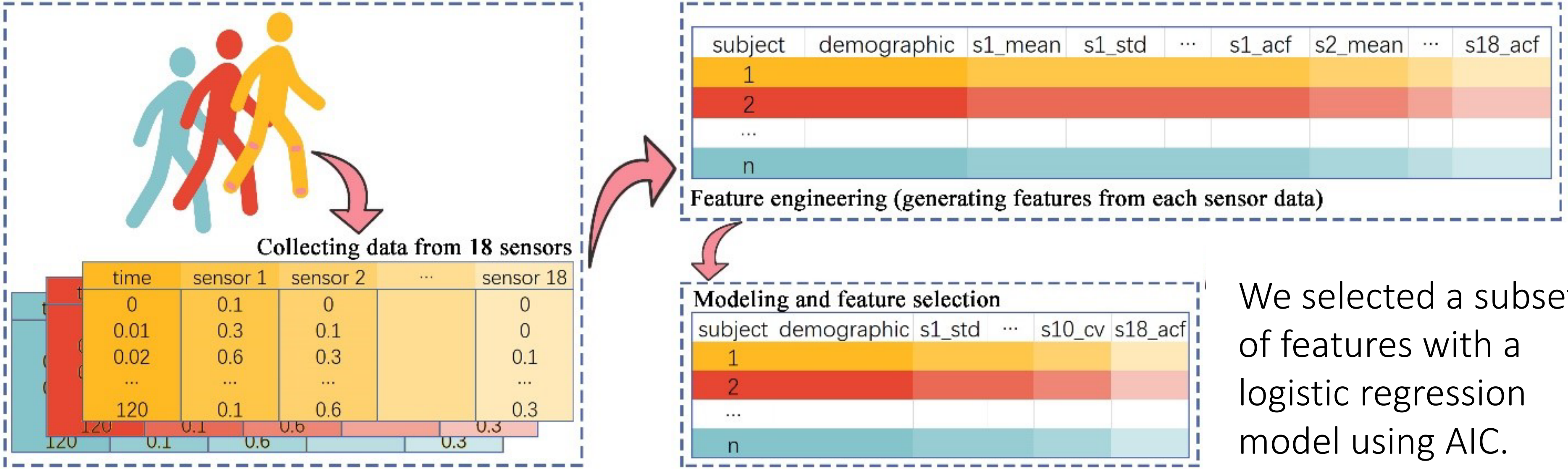
Method

Post-processing

- Some subjects had more than one records, we utilized `proc means` to aggregate the features.
- The feature set could be merged with the demographic data and then used as the primary predictor set for analysis.

Method

Feature Selection & Modeling



We selected a subset of features with a logistic regression model using AIC.

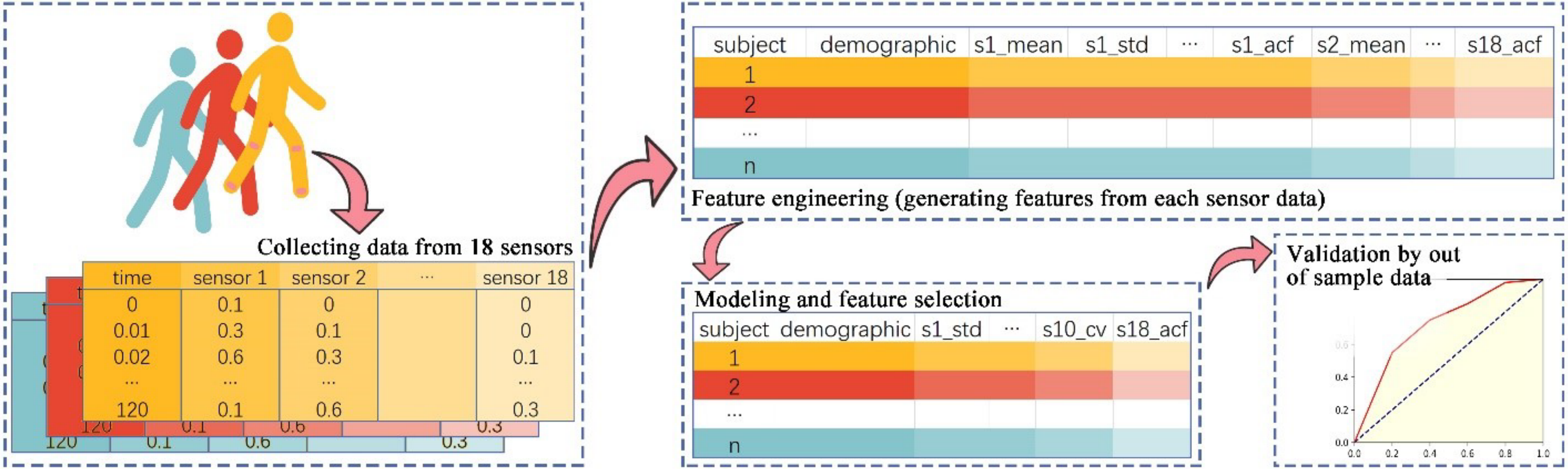
Method

Feature Selection & Modeling

- Since not all features are related to the status of disease, we used `proc glmselect` to perform effect selection in the framework of general linear models.
- Good predictors for a logistic model could be identified and selected by `proc glmselect` when fitting a binary target [Robert Cohen, 2009].
- The procedure selects a subset of features with a logistic regression model using AIC as the criterion while forcing demographic variables (age, gender, study) to be included in the model as potential confounders.

Method

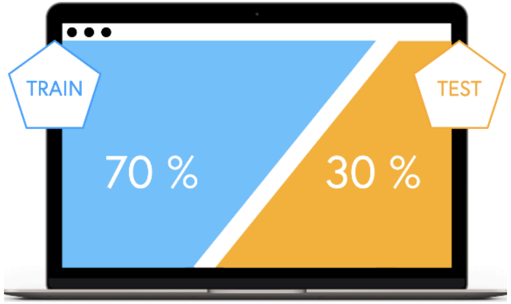
Model Validation



Finally, we validated our selected model by calculating the AUC of the testing data.

Method

Model Validation



- Specifically, we randomly separated our dataset into training data (70%) and testing data (30%) by `proc surveyselect`.
- A logistic model based on selected features was fitted on the training set using `proc logistic`, and AUC was obtained on the testing set.
- We repeated the same procedure for 100 times and got the out of sample AUCs.

Results

Feature Engineering

- 288 features in total, 18 variables and 16 features for each variable
- AIC criteria
- 7 features are selected
- Final logistic regression model includes the 7 selected features, age, gender, and study group indicator.

Results

Feature Engineering

- Odds ratio estimates and 95% confidence intervals obtained from logistic regression model.

Effect	Odds ratio	Lower 95% CI	Upper 95% CI	Pr > ChiSq
Study group 1 vs 3	1.107	0.301	4.068	**0.410**
Study group 2 vs 3	3.105	0.626	15.406	**0.122**
Age	1.032	0.965	1.103	**0.359**
Gender male vs female	5.793	1.704	19.700	**0.005**
L1 std deviation	0.963	0.943	0.984	**0.001**
L6 skewness	0.145	0.021	0.997	**0.050**
R1 range	0.994	0.987	1.001	**0.083**
R4 median	1.054	1.016	1.093	**0.005**
100x (R4 ACF 30)	1.086	1.019	1.158	**0.011**
R6 range	0.991	0.982	0.999	**0.037**
R7 CV	0.949	0.905	0.995	**0.050**

Results

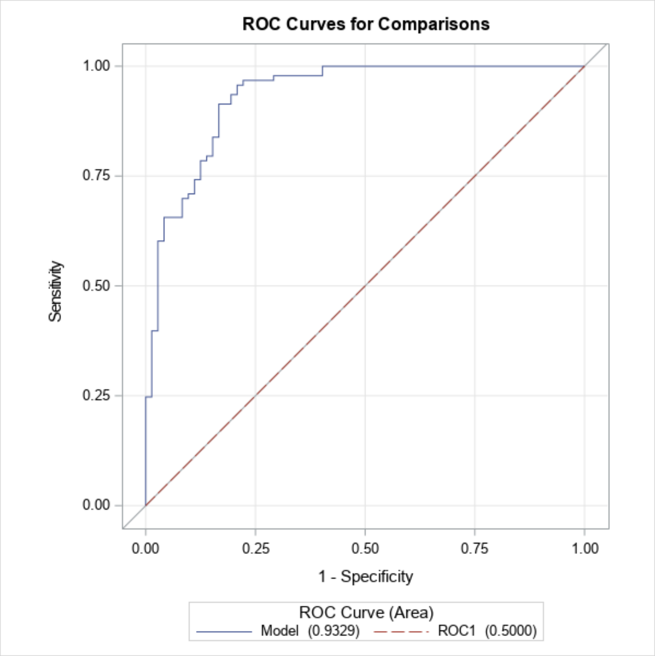
Prediction of Parkinson's Disease

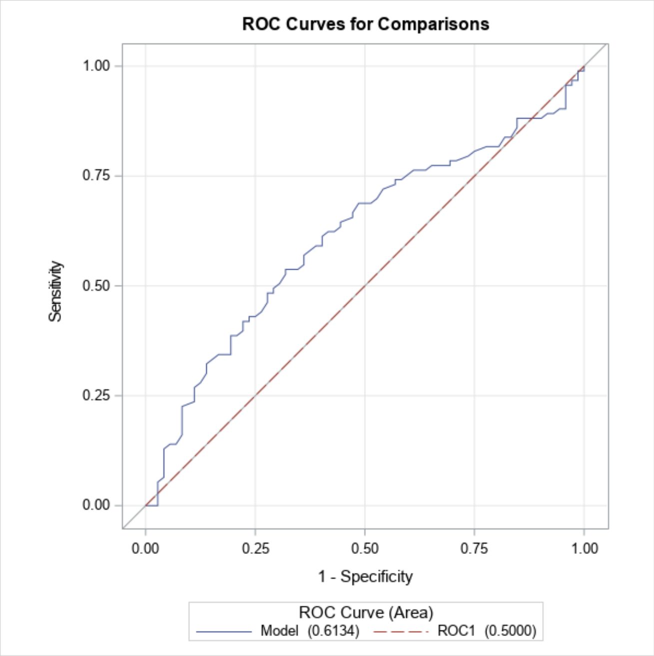
- Baseline model includes age, gender, group indicator
- Compare the final model and baseline model on the entire dataset
- ROC curve and AUC

Results

Prediction of Parkinson's Disease

- Final model (AUC=0.93), baseline model (AUC=0.61)





Results

Prediction of Parkinson's Disease

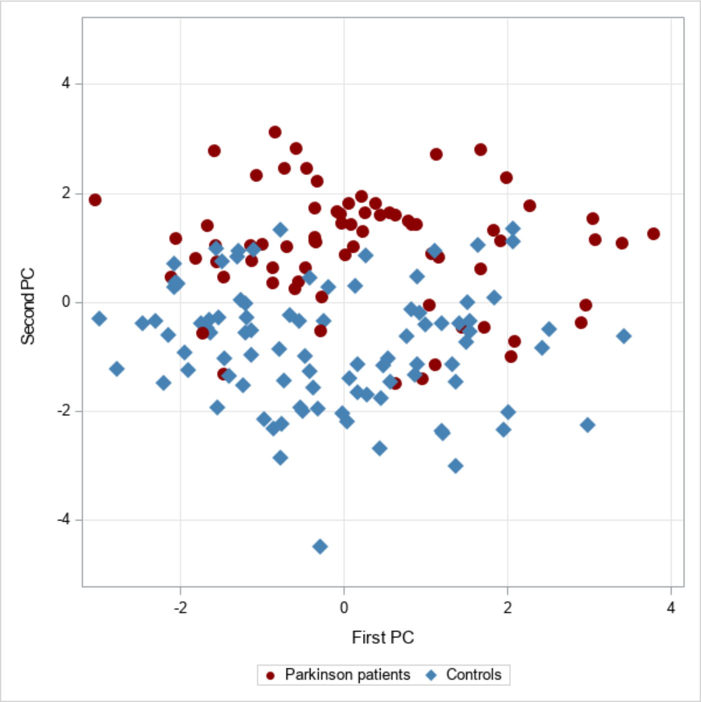
- Compare two models on test sets using a resampling method that randomly split the dataset 100 times.
- Averaged AUC on test sets

Model	Mean of AUCs	Standard deviations of AUCs
Final model	0.9039	0.0370
Baseline model	0.5303	0.0623

Results

Prediction of Parkinson's Disease

- Principal component analysis (PCA)



Inspection

Conclusion

Association

- The jitter frequency is associated with the status of PD

1) Less jitter dispersion and dispersion frequency, higher probability having PD
2) Narrower range, symmetric distribution, higher median and higher autocorrelation across time
3) PD patients have a reduced stride length and a short average swing time

➡ Walking force of PD patients is relatively stable and predictable from past records

Prediction

- A simple way to predict PD status from the walking condition of patients

1) A high predicting accuracy with AUC 0.9039 when 10 features are used

Inspection

Discussion

Advantages

- A simple method easy to understand
- Convenient to expand the application to other diseases
- High accuracy

Disadvantages

- Ignore the internal information in time series data
- More complicated features can be considered, e.g. Teager–Kaiser energy operator (TKEO), using detrended fluctuation analysis to get features

Inspection

Future work

Potential related study

- Incorporate the data from independent replicated study to do validation
- Time series analysis to fit the model
- Other advanced machine learning methods

Application

- Apply the same analysis to other rare diseases, e.g. ALS
- A method to give a whole picture of association between variables and diseases

Thank you!

Contact Information
jinge1997@outlook.com