Summer 2025 Data Science Boot Camp - Erdős Institute

Sayantan Roy Chowdhury

Min Shi

E.G.K.M. Gamlath

Characterizing Parkinson's Disease Progression: A Multifaceted Data Science Approach

A Data-Driven Investigation with machine learning and statistical analysis

Backgrounds

1976

1.6 people

Prevalence of Parkinson's disease.

1983 2.3 people 1990 3.2 people 1997 2004 6.2 people 2011 7.0 people

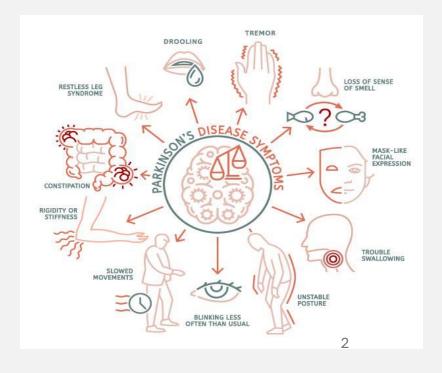
number of deaths per 100,000 people due to Parkinson's disease in the United States

> Diverse effects of Parkinson's disease

Estimated number of people with Parkinson's disease¹ per 100,000 people. Western Pacific (WHO) European Region (WHO) Doubled 2015 Data source: IHME, Global Burden of Disease (2024) OurWorldinData.org/causes-of-death | CC BY

Parkinson's disease prevalence

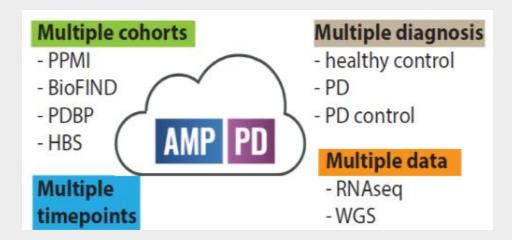
Deteriorated death risk from Parkinson's disease



Our World in Data

Data Collection

- The Accelerating Medicines Partnership Parkinson's Disease (AMP PD) dataset is a large-scale, collaborative effort collecting data for Parkinson's disease.
- Version 4.0 It brings together longitudinal clinical, genetic, and biospecimen data from 4 cohort studies: BioFIND, PPMI, PDBP, and HBS.
- The **AMP PD Version 4.0 data** was obtained by request from their knowledge platform.



Unified Parkinson's Disease Rating Scale (UPDRS)

UPDRS scores Parkinson's disease severity from 0 (healthy) to 4 (severely disabled) across 44 items in 3 key areas:



Mentation, Behavior, and Mood (4 items): Assesses cognitive and emotional health.



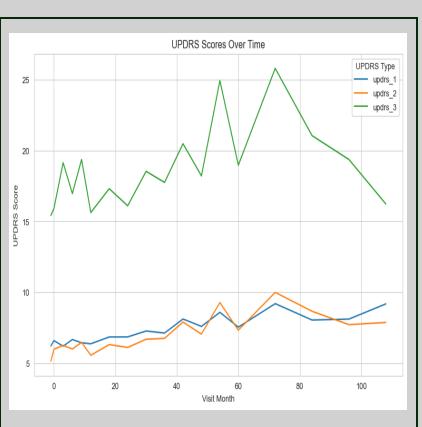
Activity of Daily Living (13 items): Measures ability to perform daily tasks independently.



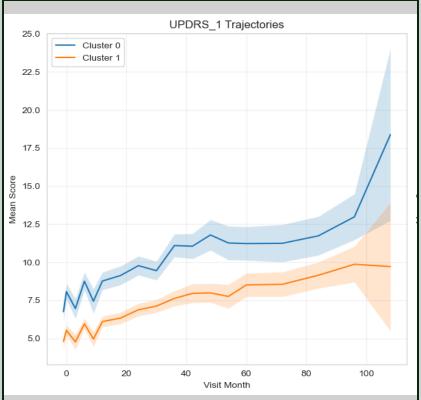
Motor Function (27 items): Evaluates muscle issues like tremor, rigidity, and bradykinesia.

The UPDRS scores are the **primary clinical assessments** for Parkinson's disease which we focus for all our analysis.

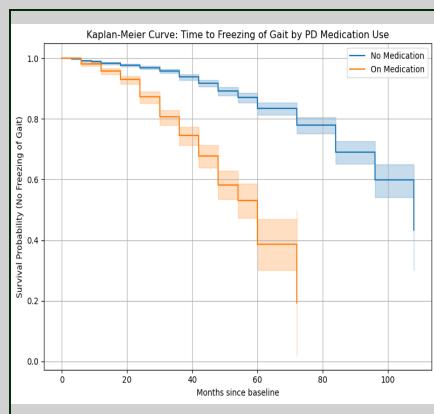
Key Questions



Question 1: Can Parkinson's scores be predicted reliably?



Question 2: Are there distinct patterns in how the disease progresses?



Question 3: What factors influence the time to freezing of gait?

Question 1: Can Parkinson's progression be reliably predicted?

-- a supervised regression task

Model variables

- Features: static and time depedent clinical results, including demographic, family history and brain scan data
- Targets: UPDRS scores, a standardized measure to assess the severity and progression of Parkinson's symptoms, as described before

Key Performance indicator:

- Mean absolute error, measuring the deviation of prediction from the true target values.
- The lower the MAE, the better the performance.

Stakeholders:

Parkinson patients & Healthcare providers

Data preparation and Preprocessing

Final Data

set

(Over 20000

observation

s from 4000

partipants,

with

about 1400

features)

Raw data

Demogr. Data (Age, Sex, Race)

Caffeine Usage History

Longitudinal Diagnosis Data

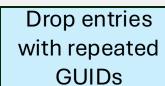
Family History

Biospecimen tests (1345 features)

Brain Imaging tests

Operations

Remove redundant columns



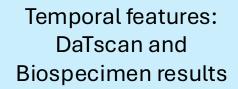
Handling change in the report

Pivotting & Averaging & Dropping results with LOG

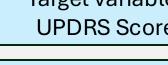
Patient-disjoint Train-Test Split

Training data:

Static features: Caffeine Usage, Age, Race, Sex, Family History and DTI Scan results

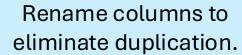


Target variables: **UPDRS Scores**





Identify rare categories (from training data), group them, and apply onehot encoding.



Drop features with 100% missing data.



Test Data (Untouched Until Final Stage)



The Challenge: Little Predictive Power

Low correlations among features

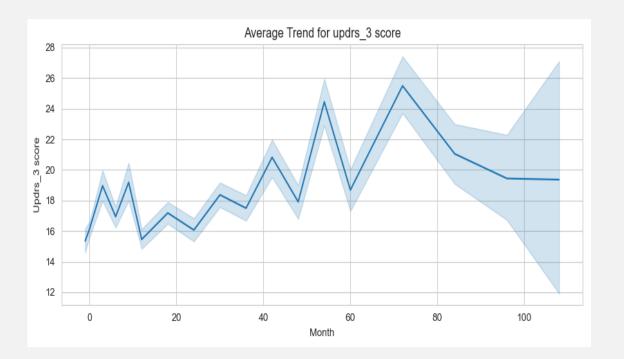
• Little predictive power

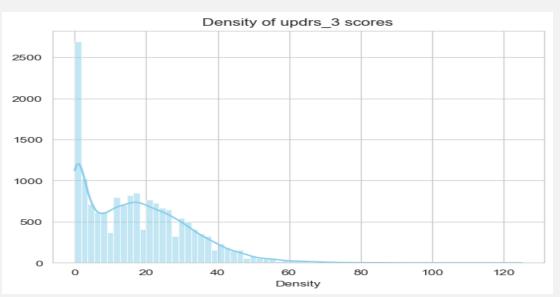
	Correlation of Features with UPDRS Scores				
age_at_baseline	0.11	0.09	0.15		
visit_month	0.09	0.10	0.09		
Greater than 16 years	-0.02	-0.06	-0.06		- 0.10
Less than 12 years	0.03	0.03	-0.01		
Other_education_level_years	0.01	-0.02	-0.01		
Unknown	-0.02	0.03	0.08		- 0.05
Male	-0.02	0.12	0.12		
Multiracial	-0.03	-0.03	-0.05		
Other_race	0.05	0.05	0.02		- 0.00
White	-0.01	-0.01	0.01		
caff_drinks_current_use	0.01	-0.06	-0.01		
caff_drinks_ever_used_regularly	0.05	-0.02	0.04		0.05
biological_mother_with_pd	-0.01	-0.06	-0.09		
biological_father_with_pd	-0.05	-0.11	-0.12		
other_relative_with_pd	0.02	-0.03	-0.07		0.10
	updrs_1	updrs_2	updrs_3	7	

Challenge Continued: Irregular Data Distribution

- Non-linear Trend
- Significant heterogeneity
- Skewed distribution of scores

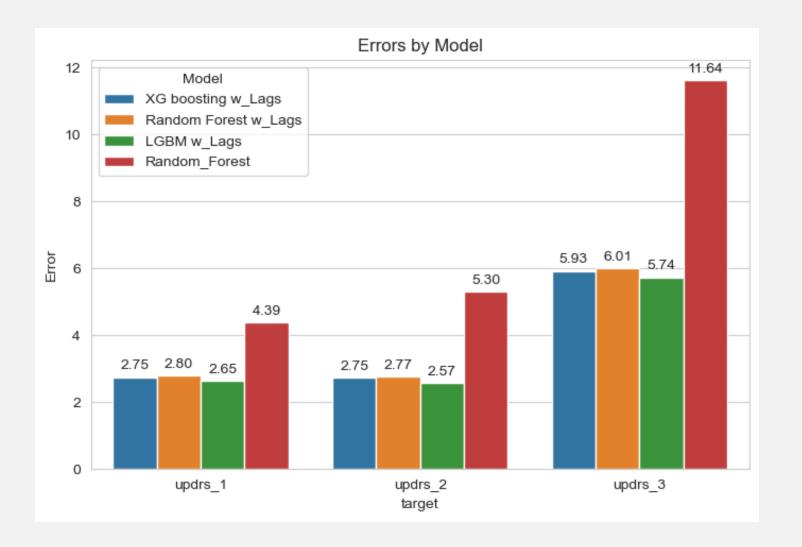
Conclusion: Together with the low correlation challenge, we decide to use tree models.





The nature of time-series data

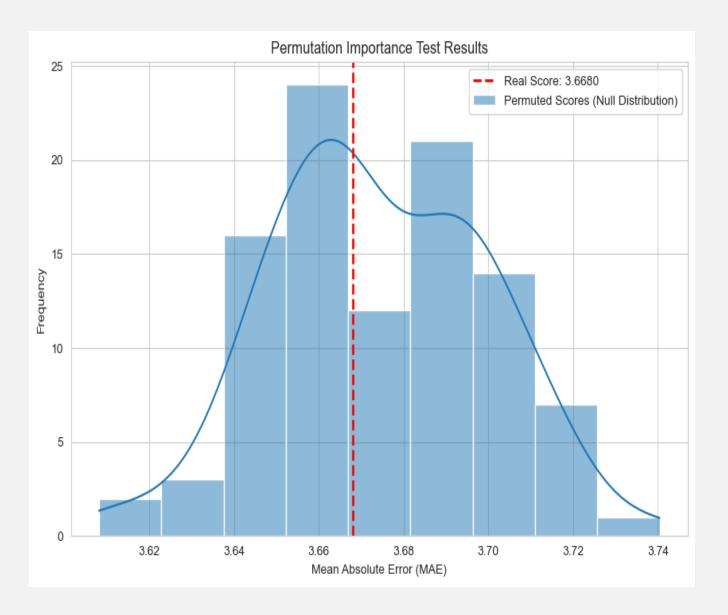
- The hidden signal: Past information
- Action: Create lagged features and re-employ tree models
- Results: Much better errors!



Will including bio-specimen data help more?

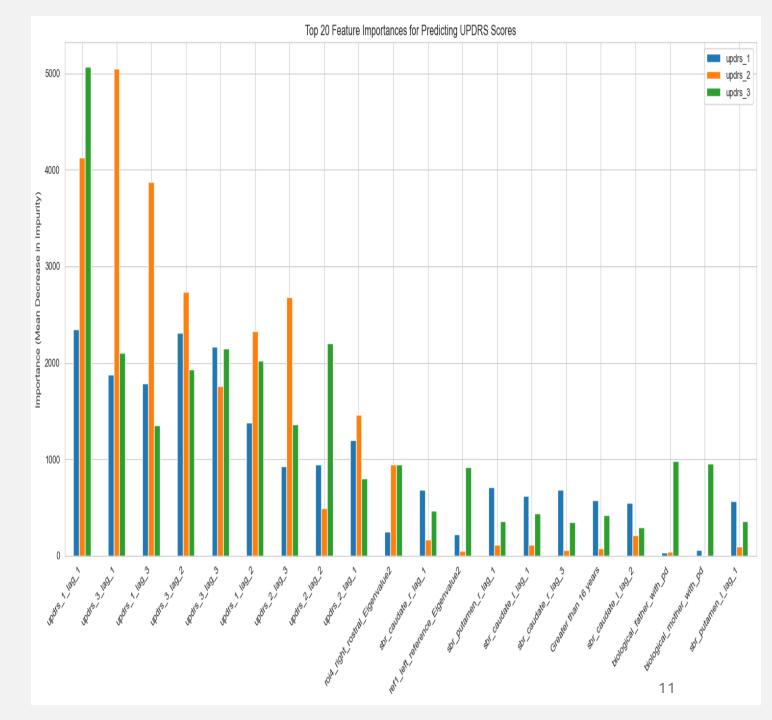
Null Hypothesis: The biospecimen data contain little signal.

Hard to obtain signals from the biospecimen data.



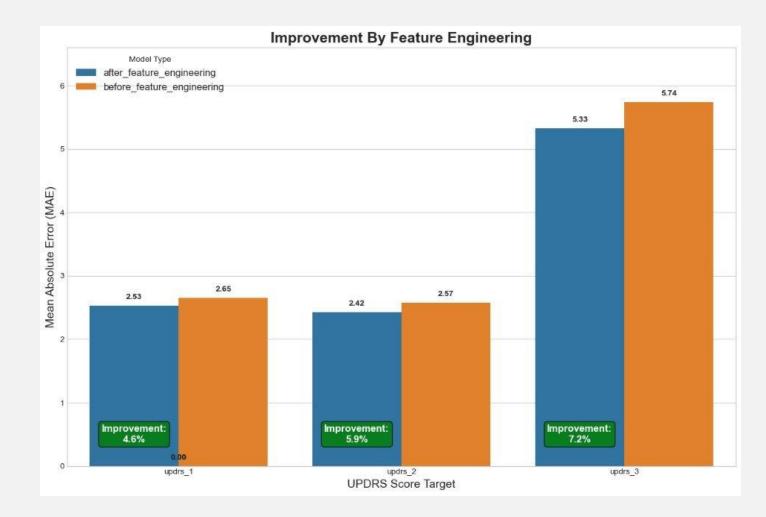
Can we further improve the model?

 Feature importance points to the direction of feature engineering



Can we further improve the model?

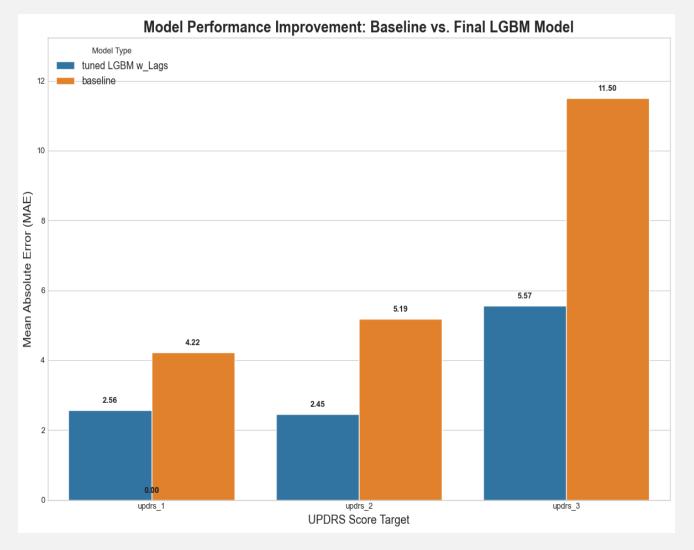
- Action: Created more timeaware features for target change.
- Result: The model is further improved!



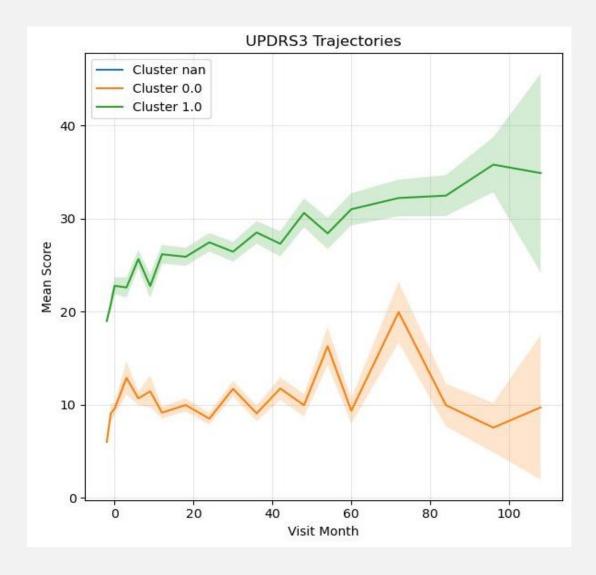
Final Pipeline & Metric

It is indeed True that Parkinson's progression be reliably predicted!

Function transformer (preprocessing & creating lags) Imputer (MICE imputer) LightGBM Model (With tuned hyperparameters)



Question 2: Can clustering capture different progression trends?



Clustering Progression Trajectories

Raw data

Operations

Features

Clustering and Scores

Longitudinal Diagnosis Data



Clinical Enrolment Dara



- Previous preprocessing.
- Use data from baseline through month
 24 only
- Resricting to data that has atleast 3 minimum visits
 - Filter out the Parkinson's Disease Patients using the Clinical Enolment Data,



Extracted Features for clustering

- Mean
- Median
- Standard deviation
 - Range
 - Slope



Clustering using 4 algorithms

- Gaussian Mixture Model
- K Means Clustering
 - Hierarchial Clustering
 - DBScan

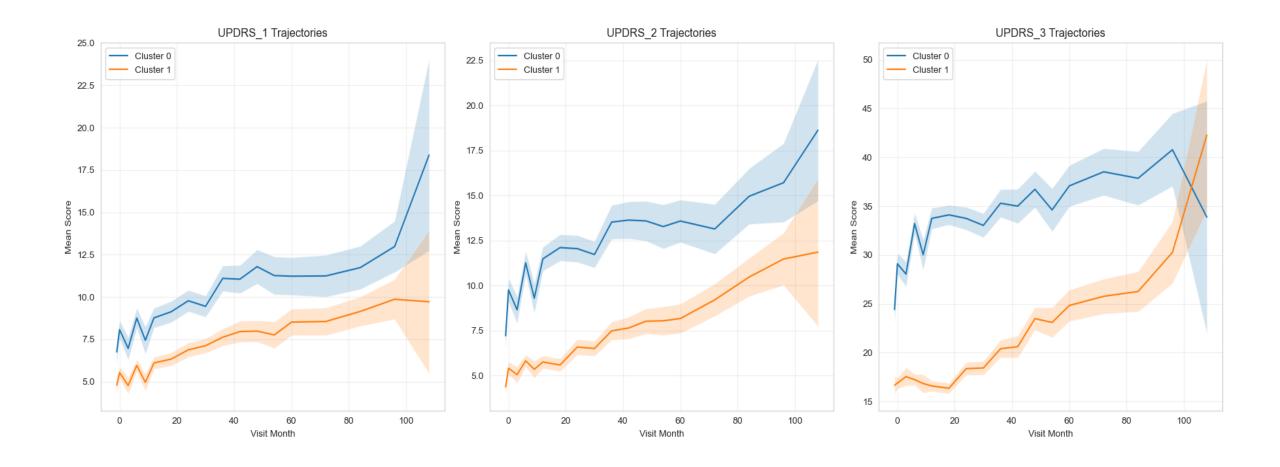
Number Of Clusters



Evaluation

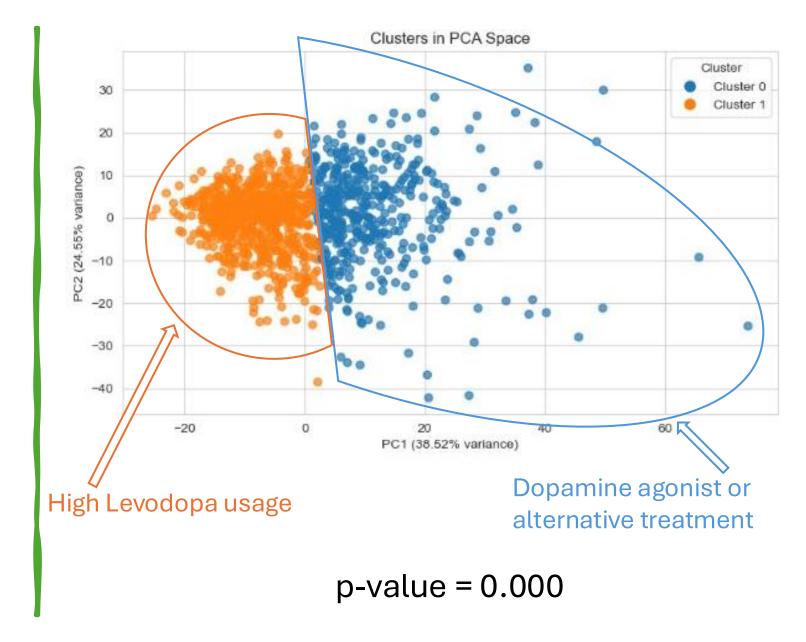
Scored using

- Silhoutte Score
- Calinski-Harabasz
 Score
- Davies-Bouldin Score



Cluster Trajectories For Each Scores By K Means (Cluster = 2)

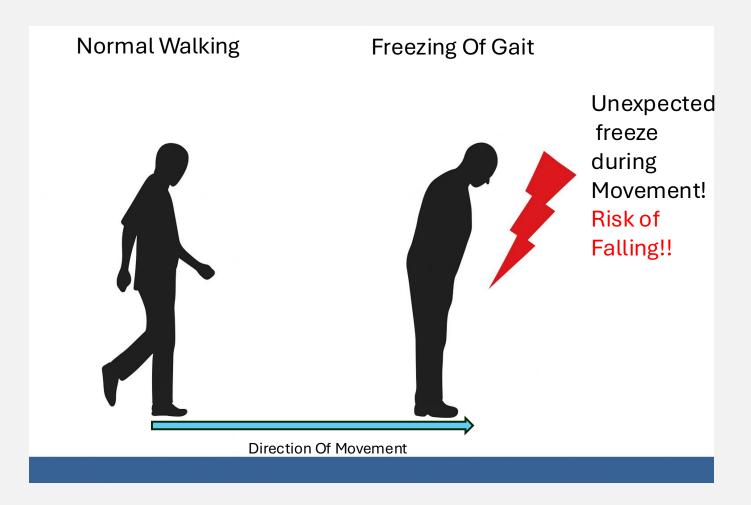
Do The Clusters have Biological Significance?



Freezing of Gait is a serious symptom of Parkinson's Disease.

Question 3:
What factors
influence
the time to
freezing of gait

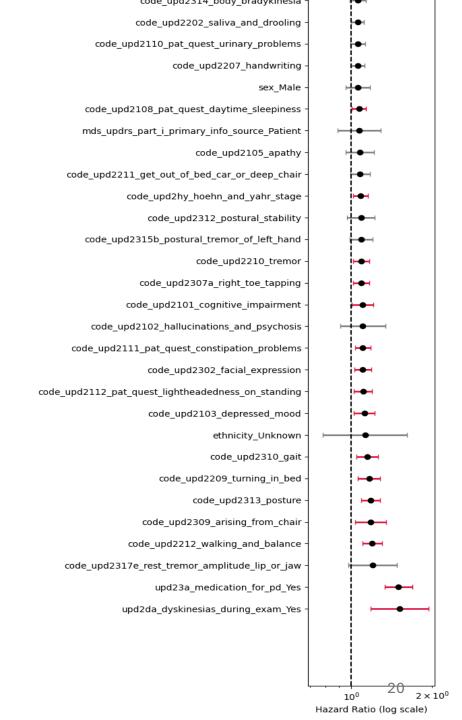




Definition: UPDRS Part 2 Q13 > 0 or UPDRS Part 3 Q11 > 0

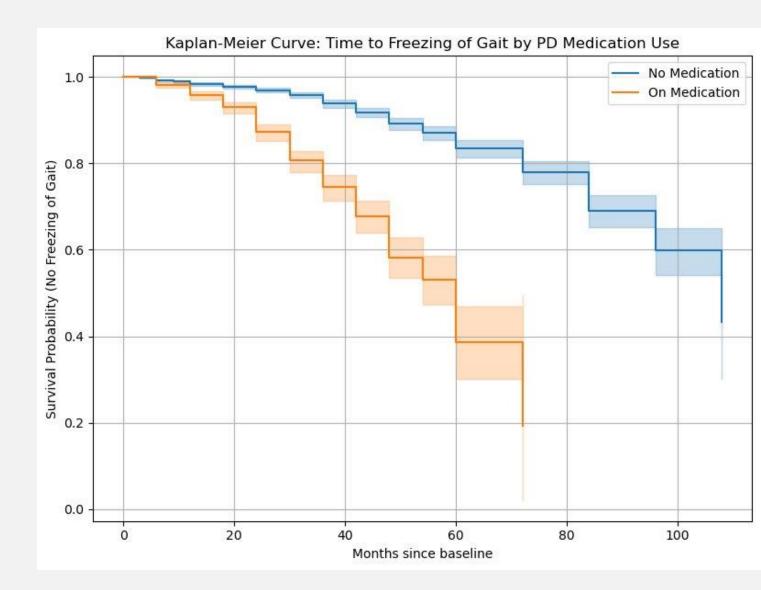
Cox Proportional Hazard for FOG using Baseline Data

- Black dots: Estimated hazard ratio
- Horizontal lines: 95% confidence interval (CI) around each hazard ratio.
- The vertical dashed line: no effect.
- Red lines: covariates that are statistically significant (p < 0.05).
- Gray lines : covariates that are not statistically significant.



Kaplan Meier Curve

- Shows the survival probability when a condition is satisfied.
- The most significant covariate:
 PD medication used!!!
- Currently validated in the research literature.







Thank You!!!