Summer 2025 Data Science Boot Camp - Erdős Institute

Sayantan Roy Chowdhury

Min Shi

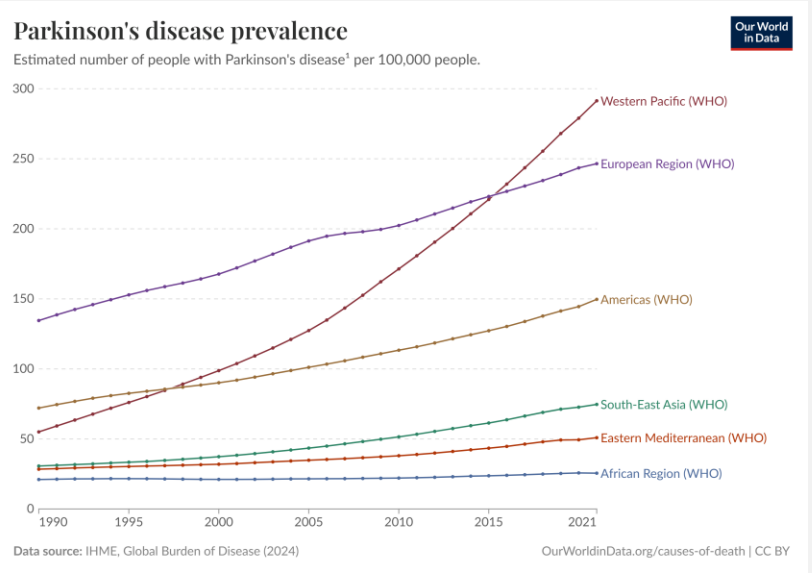
E.G.K.M. Gamlath

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

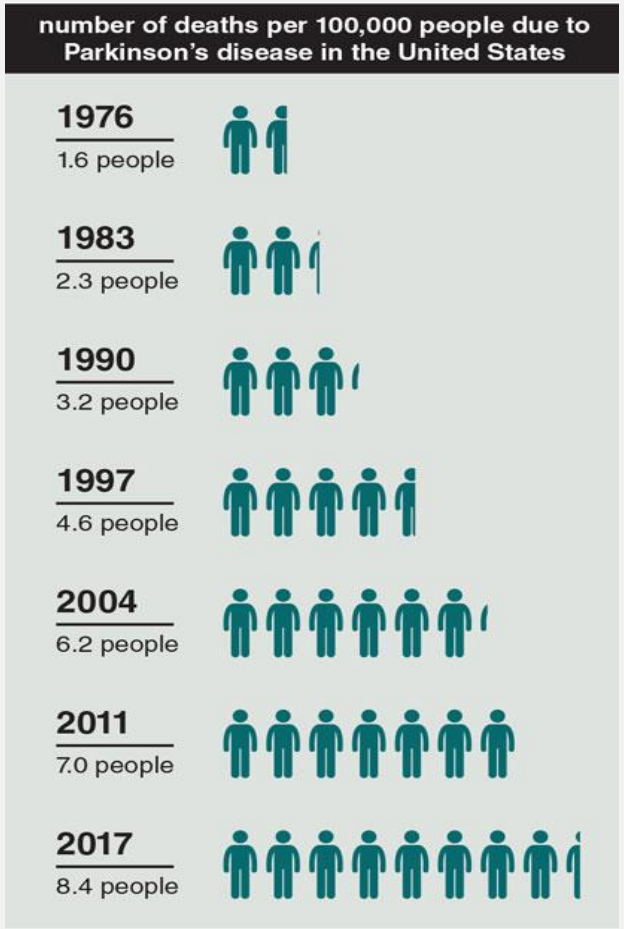
A Data-Driven Investigation with machine learning and statistical analysis

Backgrounds

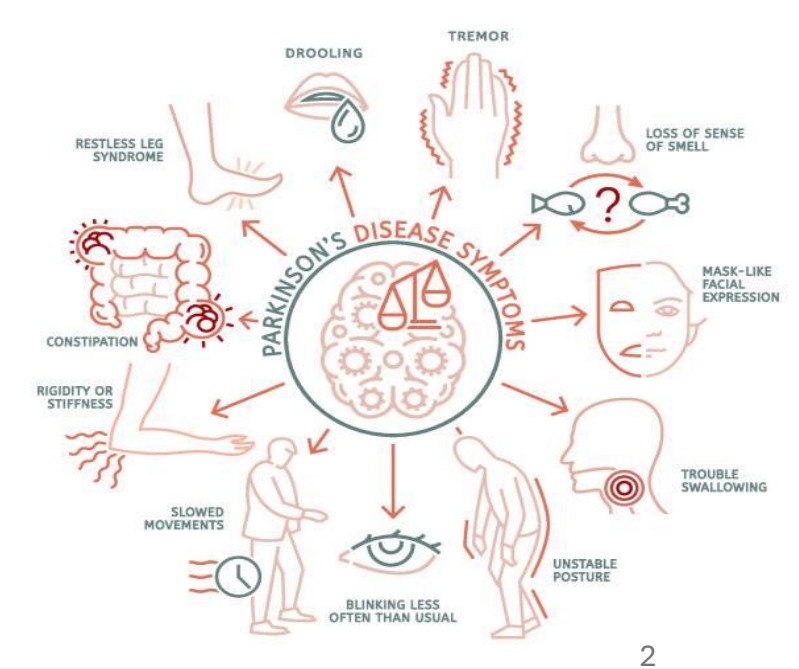
Doubled Prevalence of Parkinson's disease.



Deteriorated death risk from Parkinson's disease

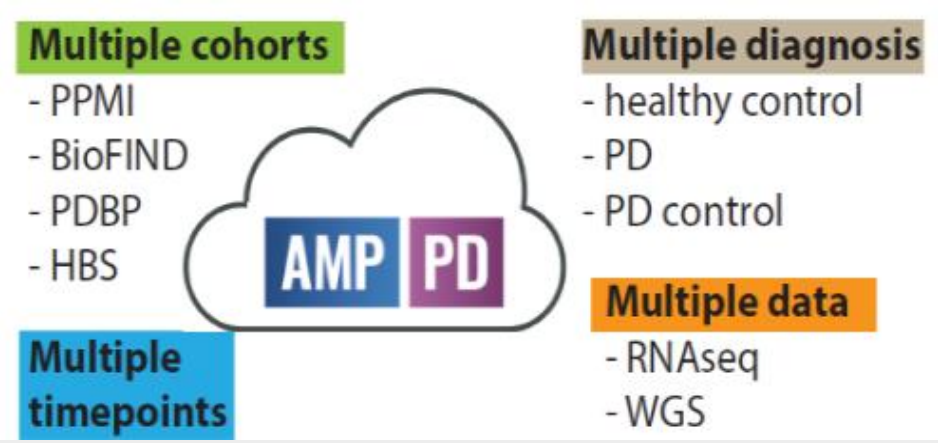


Diverse effects of Parkinson's disease



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.

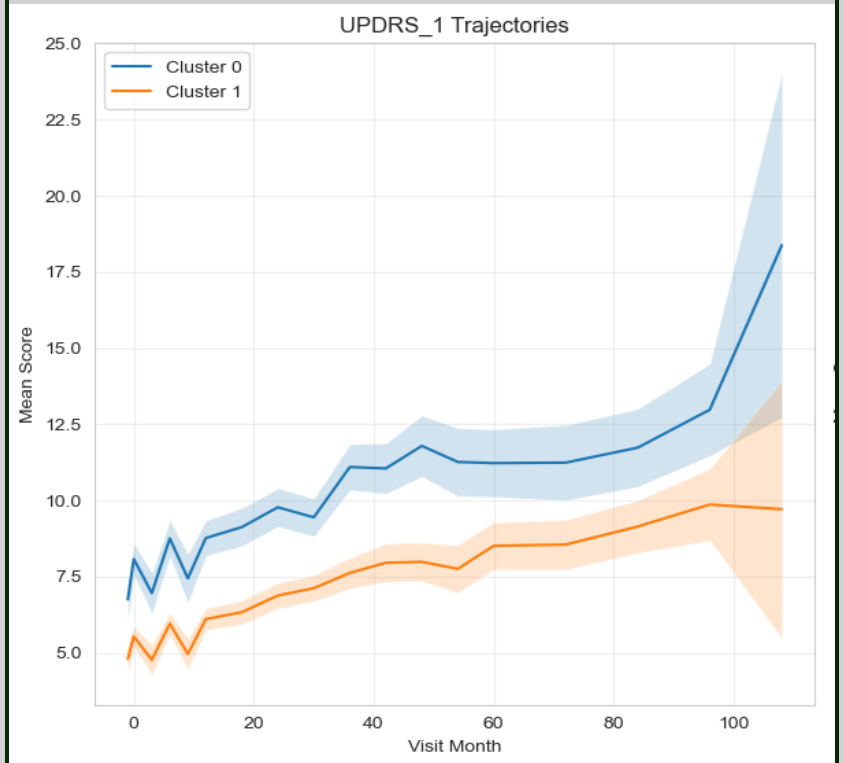
Activity of Daily Living (13 items): Measures ability to perform daily tasks independently.

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

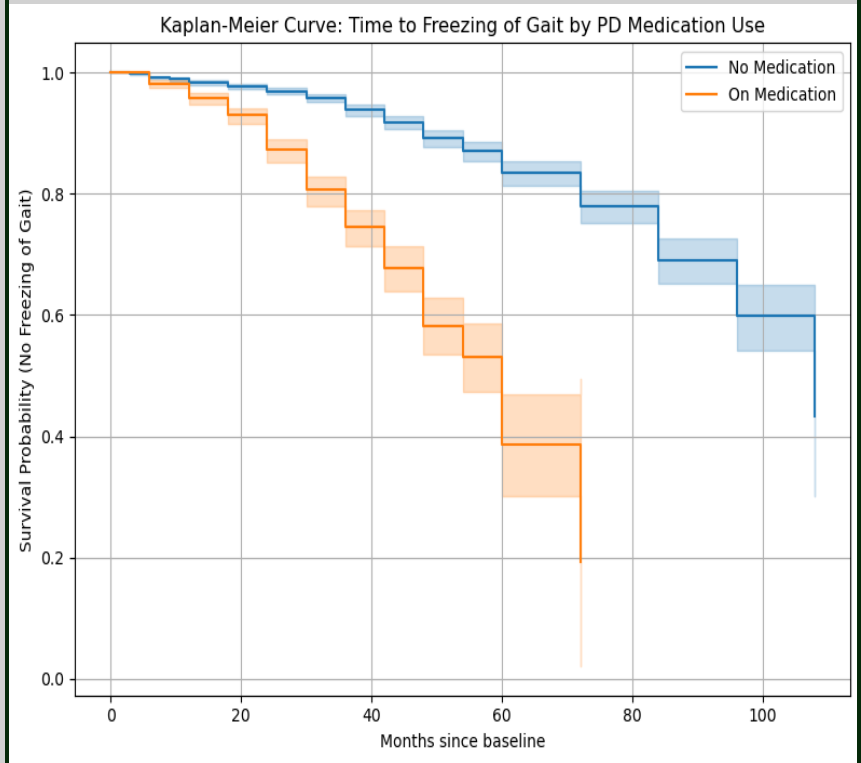
The UPDRS scores are the **primary clinical assessments** for Parkinson's disease which we focus for all our analysis.

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

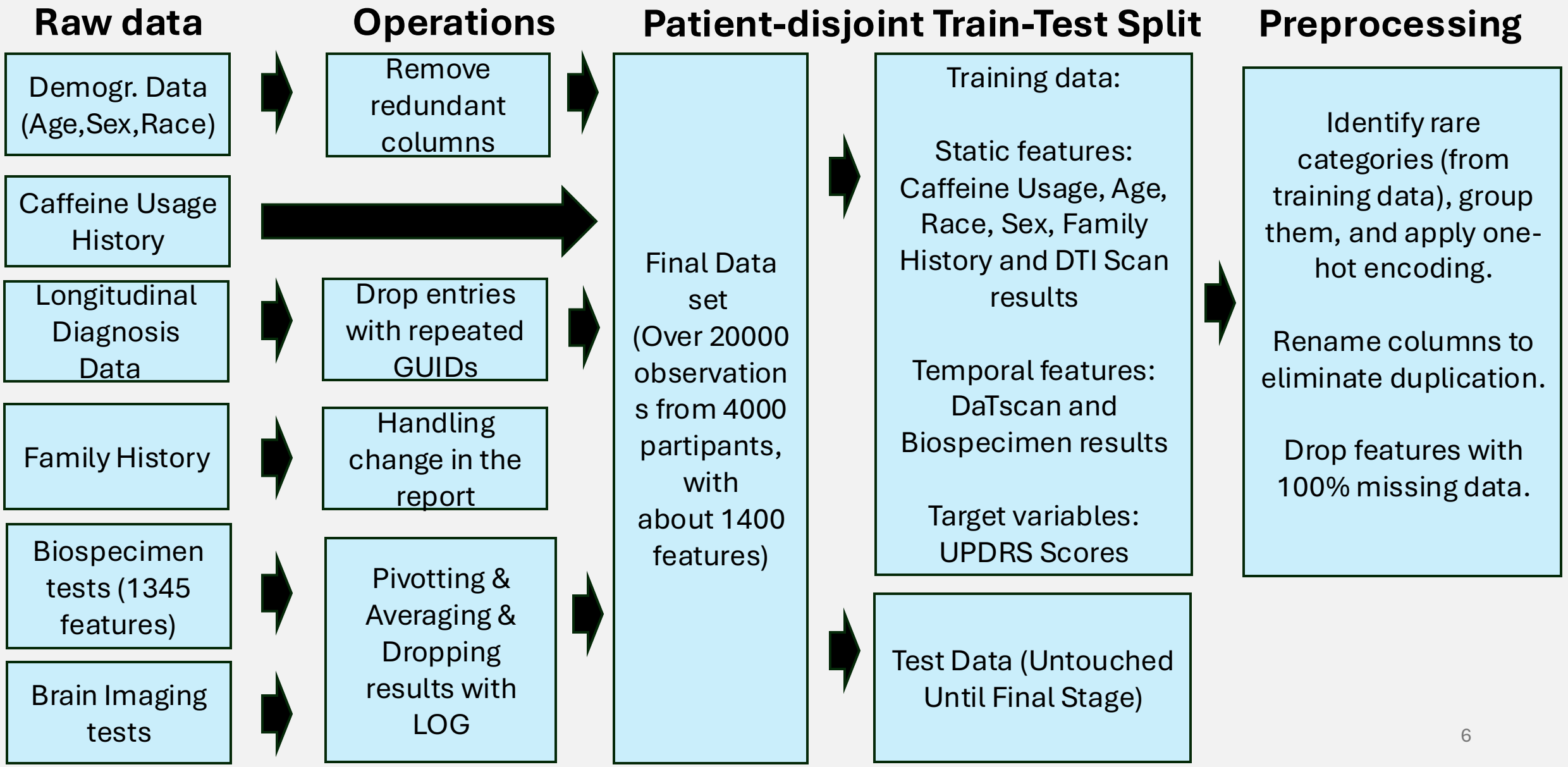
Key Performance indicator:

- 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



The Challenge: Little Predictive Power

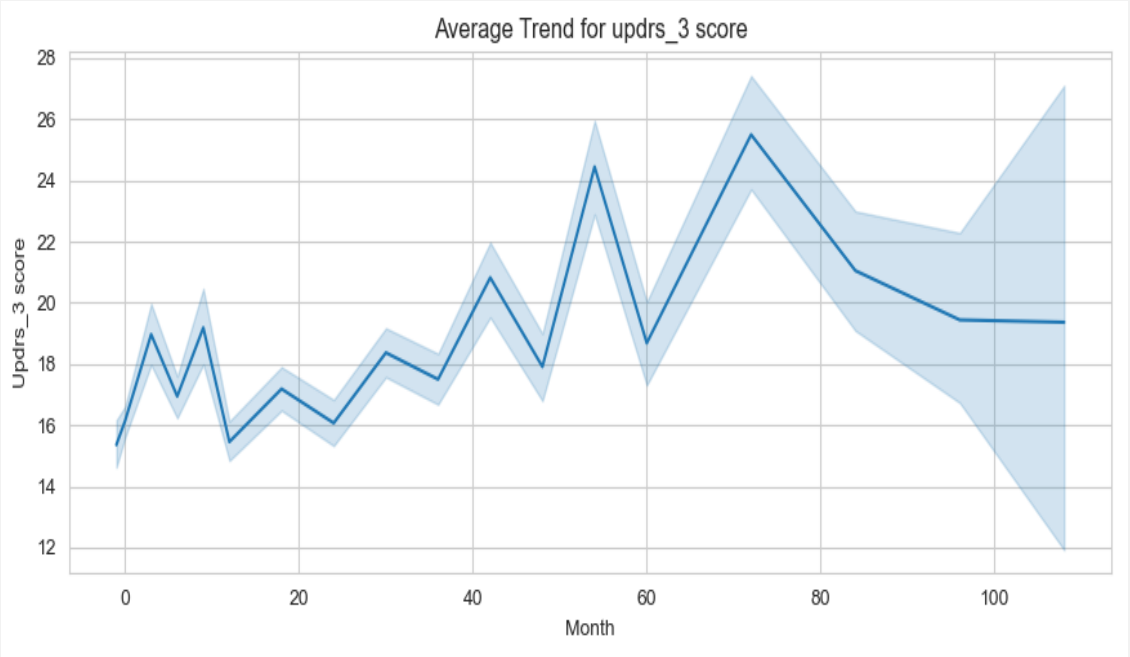
- Low correlations among features

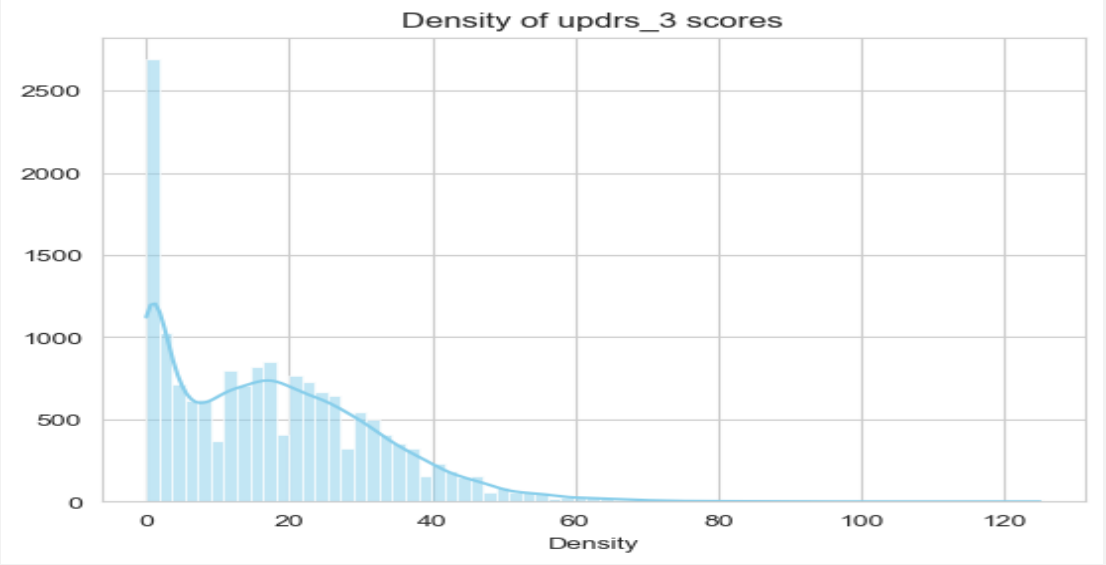
- Little predictive power

Challenge Continued: Irregular Data Distribution

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

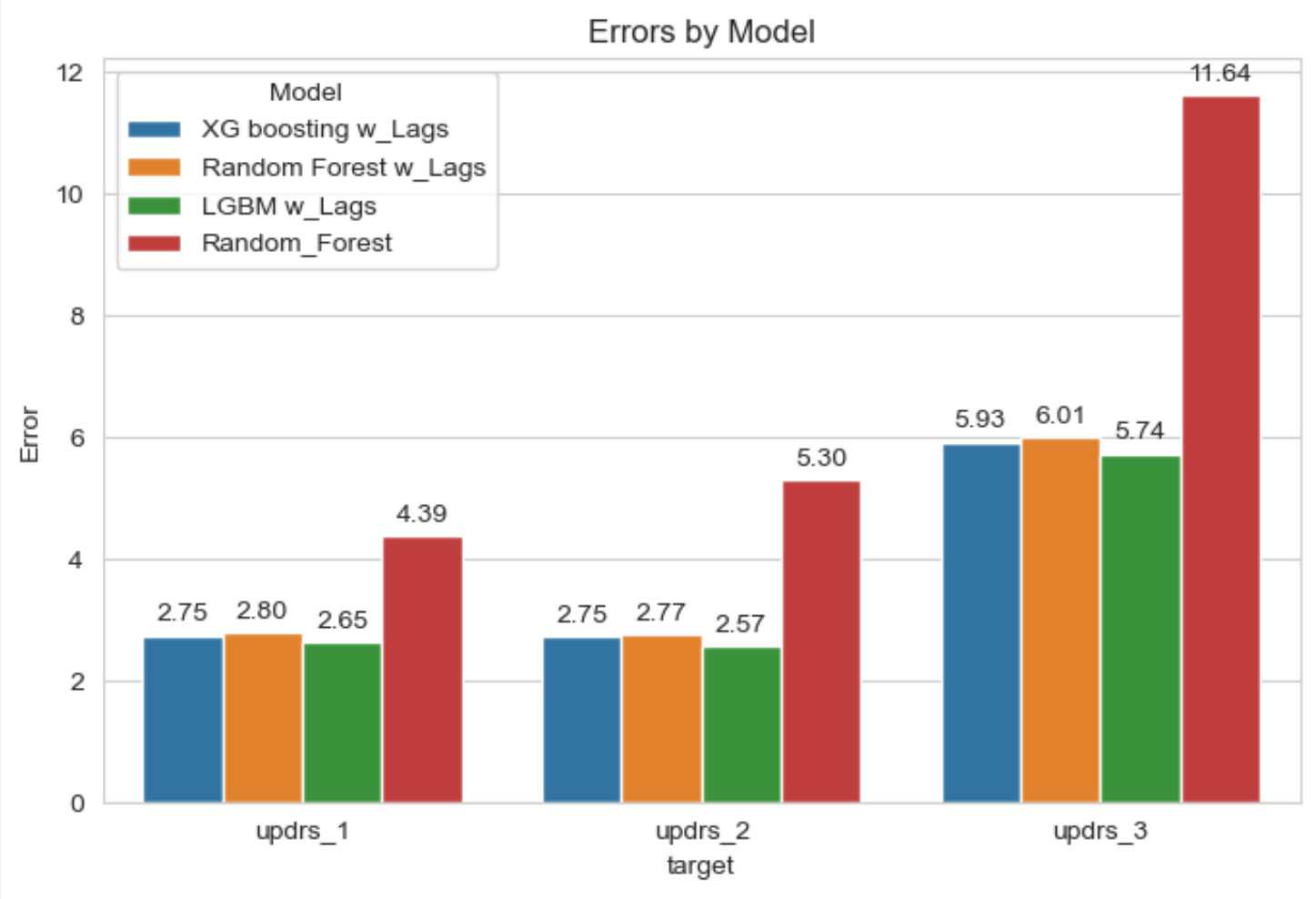
Conclusion: Together with the low correlation challenge, we decide to use tree models.





The nature of time-series data

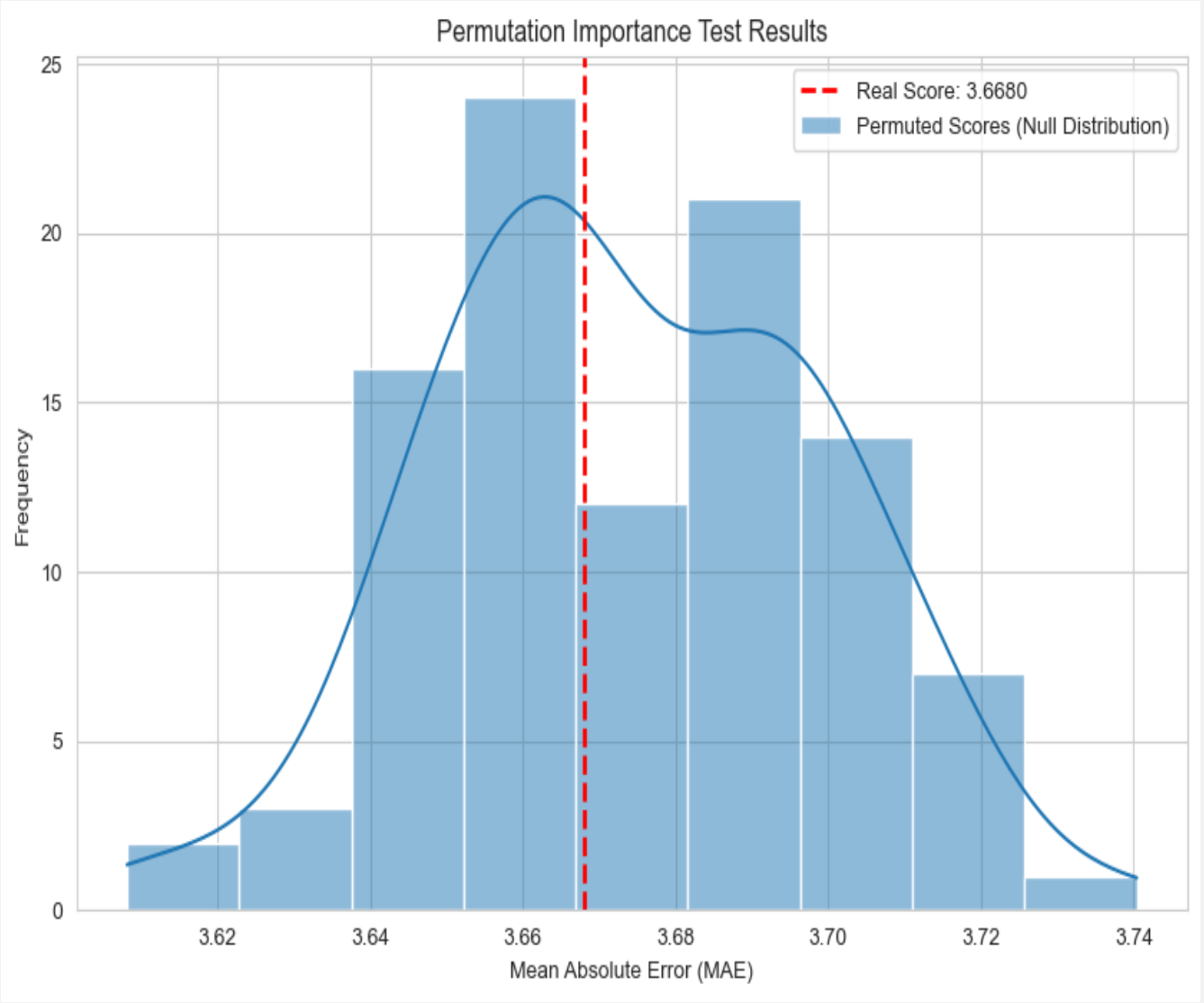
- The hidden signal: Past information
- Action: Create lagged features and re-employ tree models
- Results: Much better errors!



Will including bio-specimen data help more?

Null Hypothesis: The biospecimen data contain little signal.

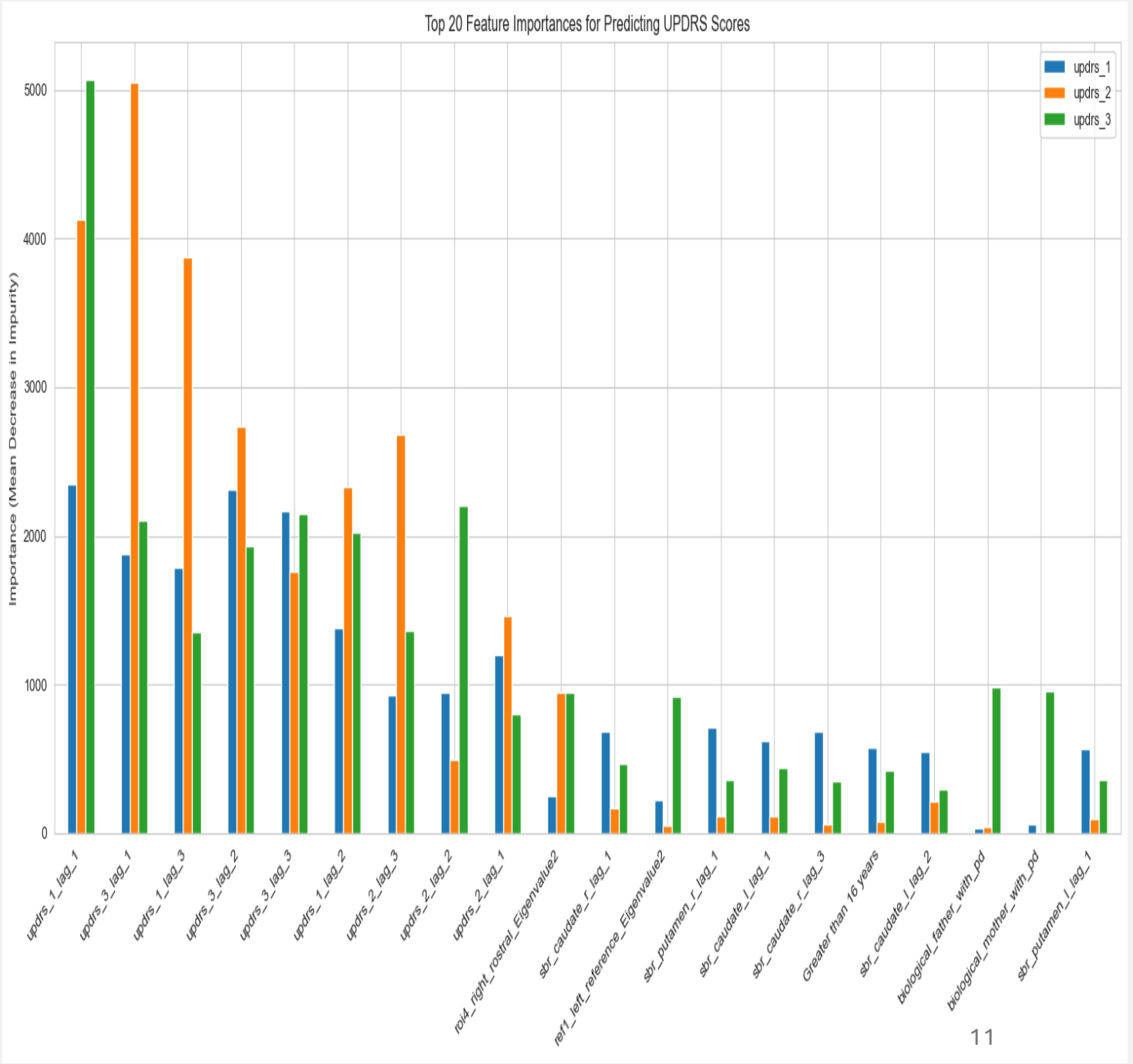
```
================================================
--- Hypothesis Test Conclusion ---
Real Model Score (MAE): 3.6680
Mean Permuted Score (MAE): 3.6742
P-value: 0.4600
================================================
```



Hard to obtain signals from the biospecimen data.

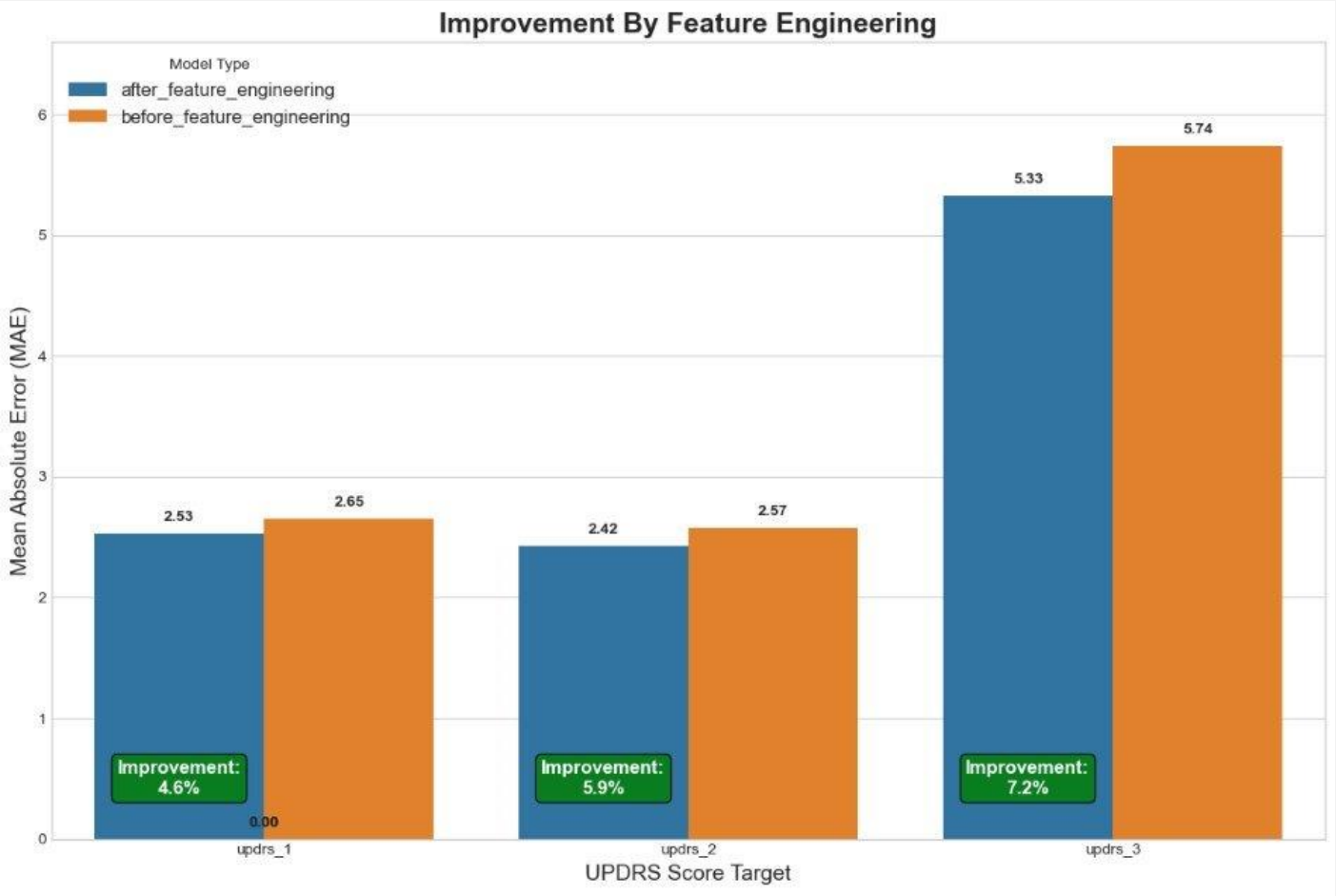
Can we further improve the model?

- Feature importance points to the direction of feature engineering



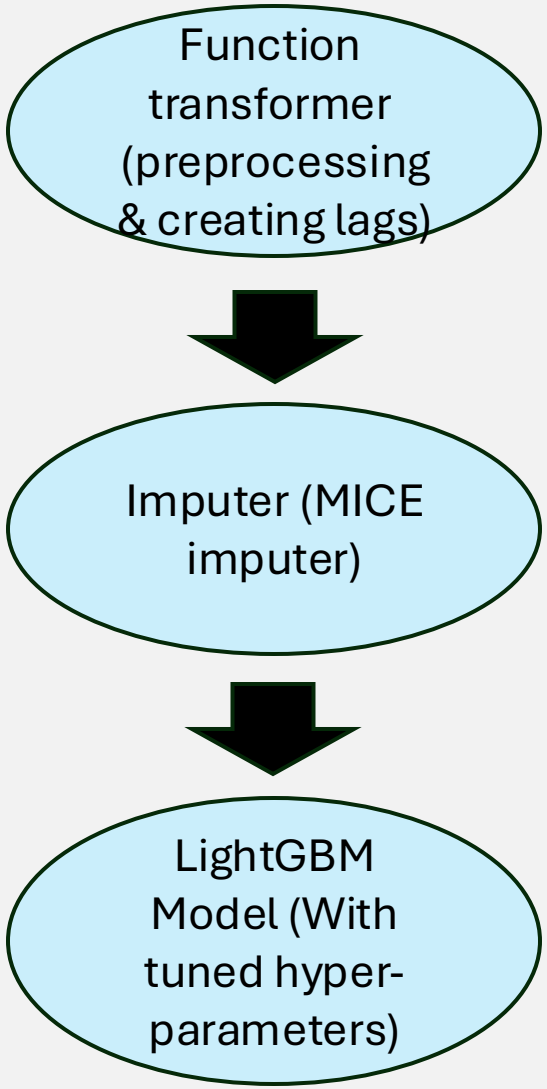
Can we further improve the model?

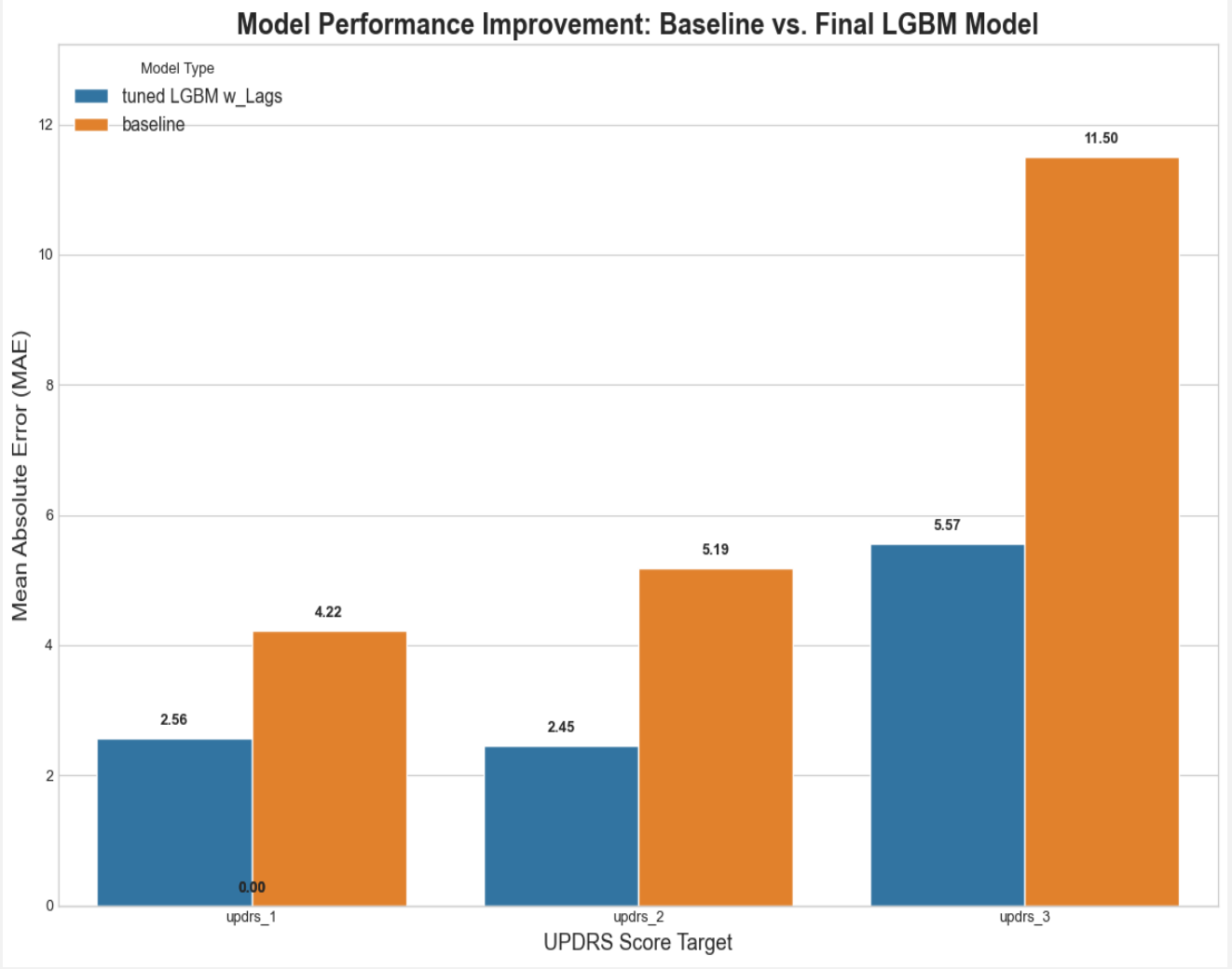
- Action: Created more time-aware features for target change.
- Result: The model is further improved!



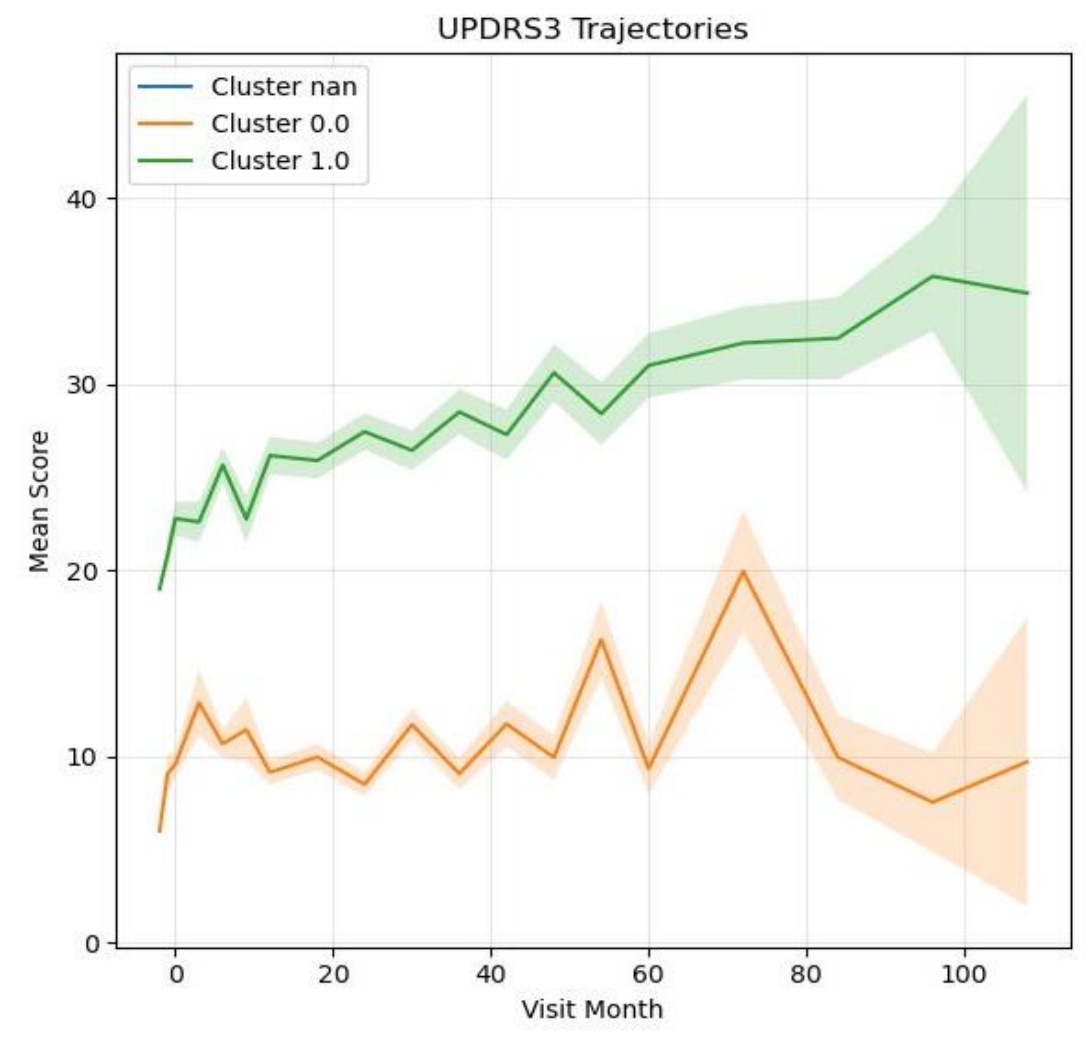
Final Pipeline & Metric

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

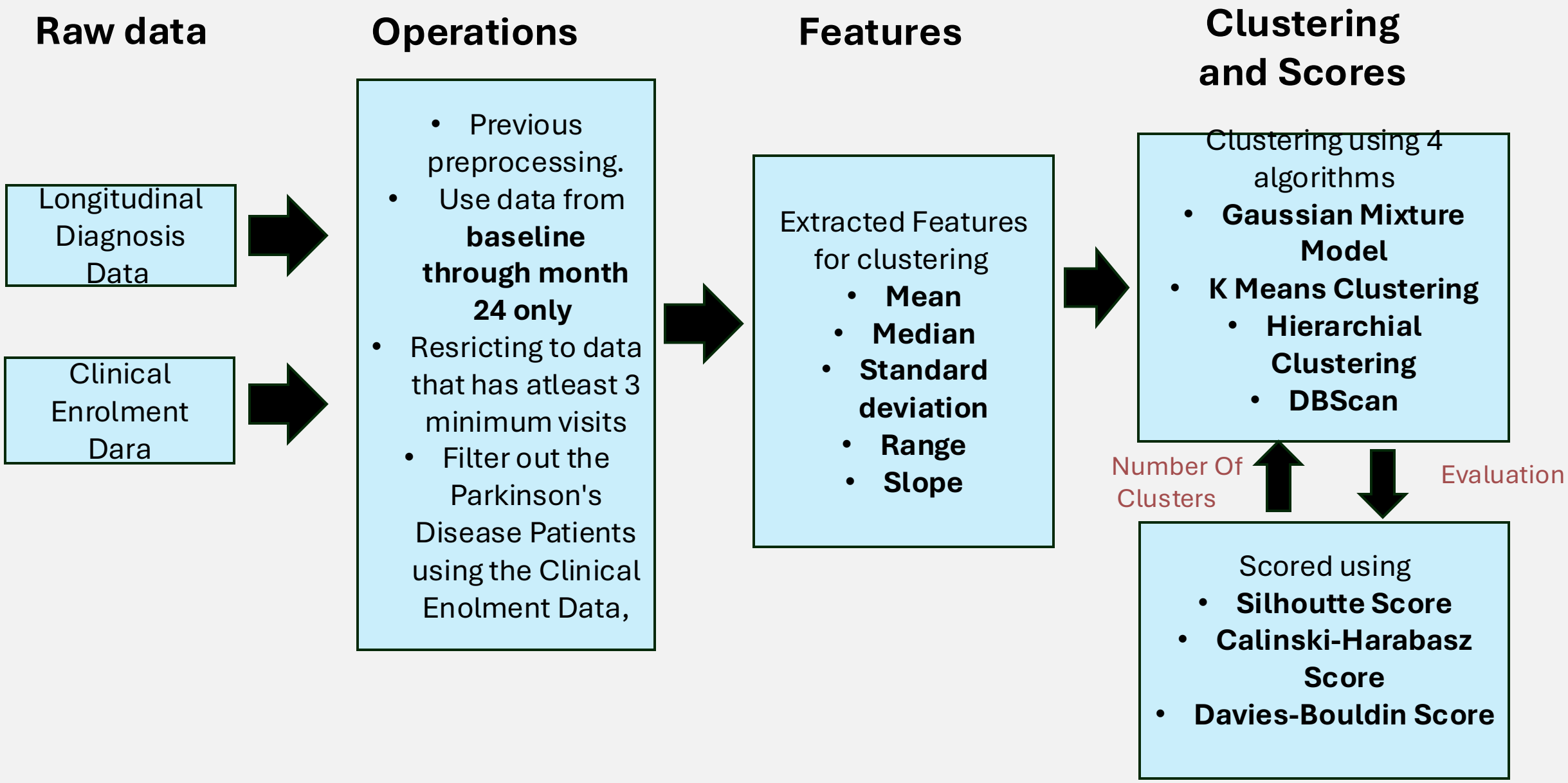


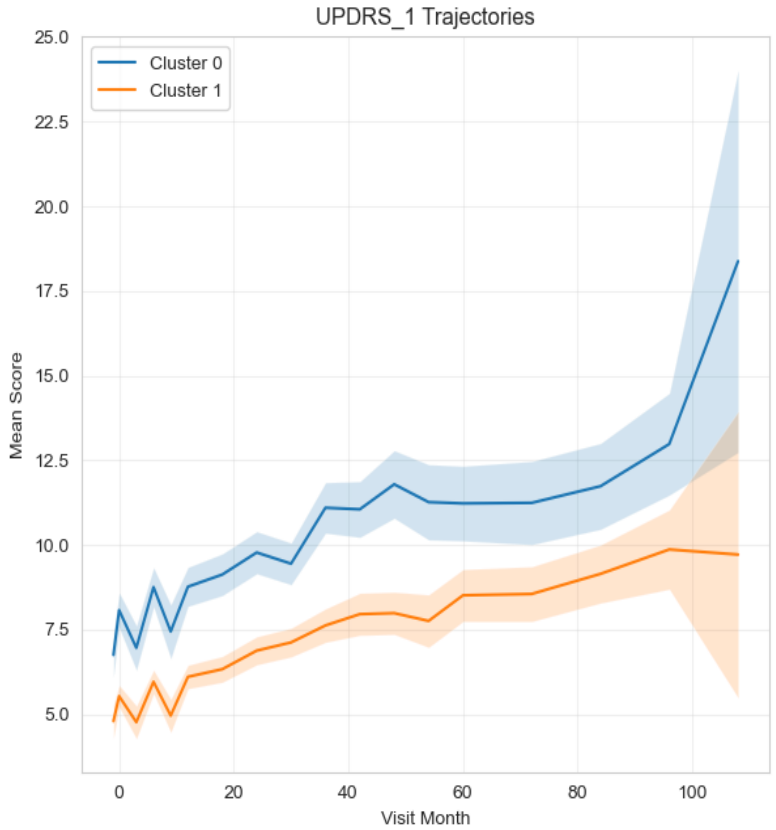


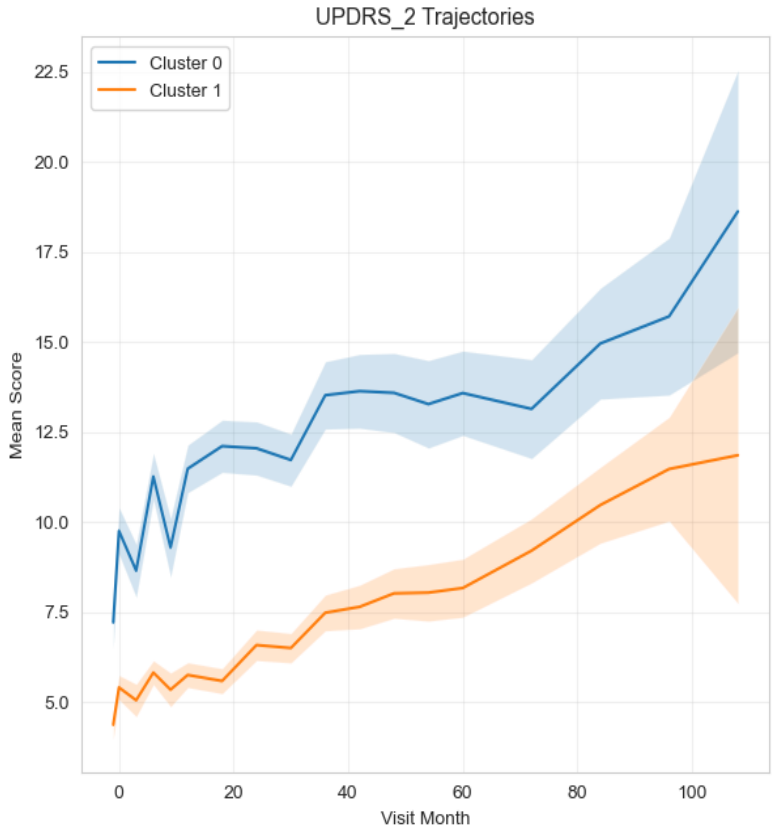
Question 2: Can clustering capture different progression trends?

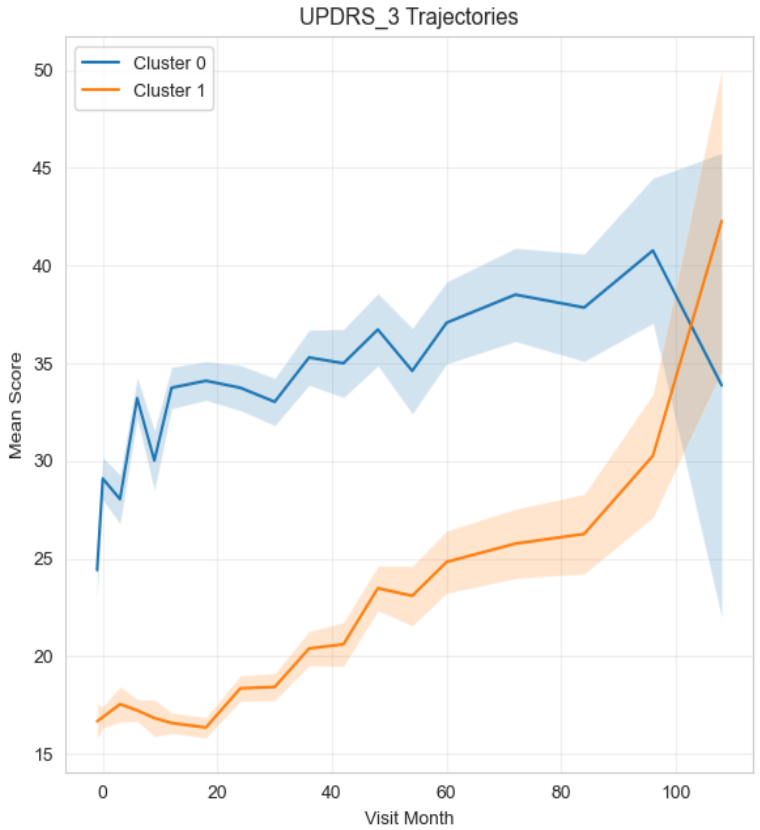


Clustering Progression Trajectories



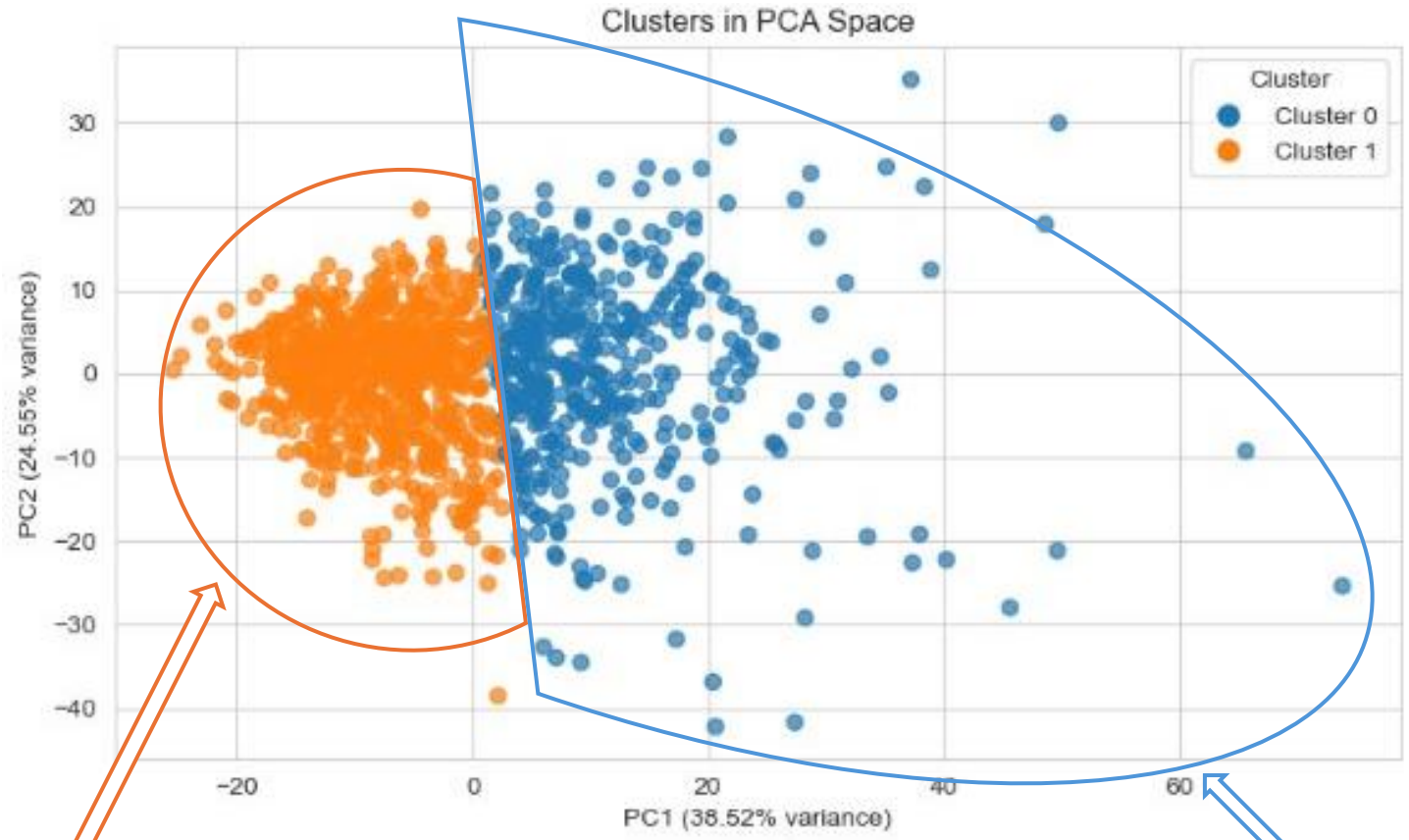






Cluster Trajectories For Each Scores By K Means (Cluster = 2)

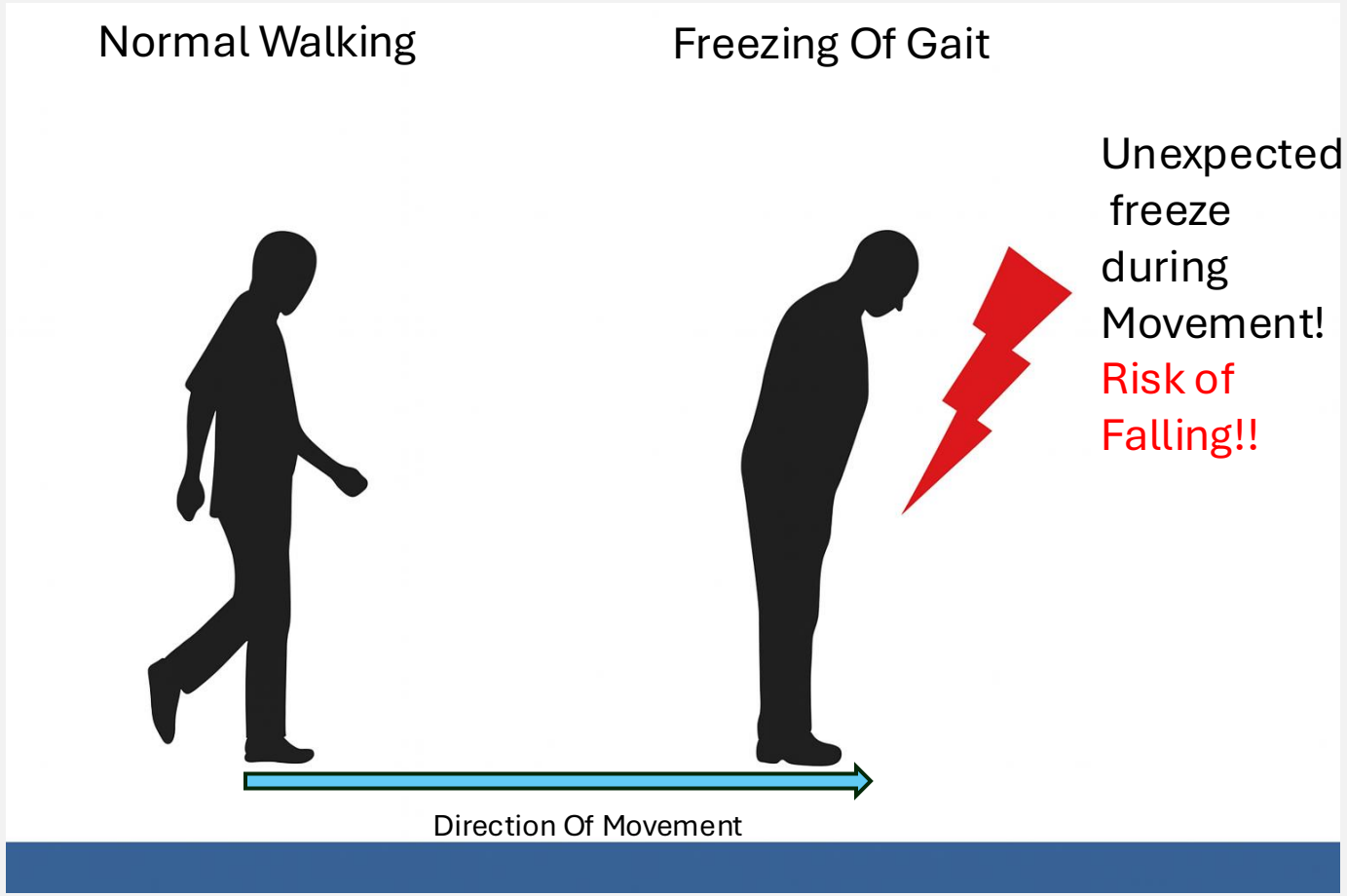
Do The Clusters have Biological Significance?



p-value = 0.000

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

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



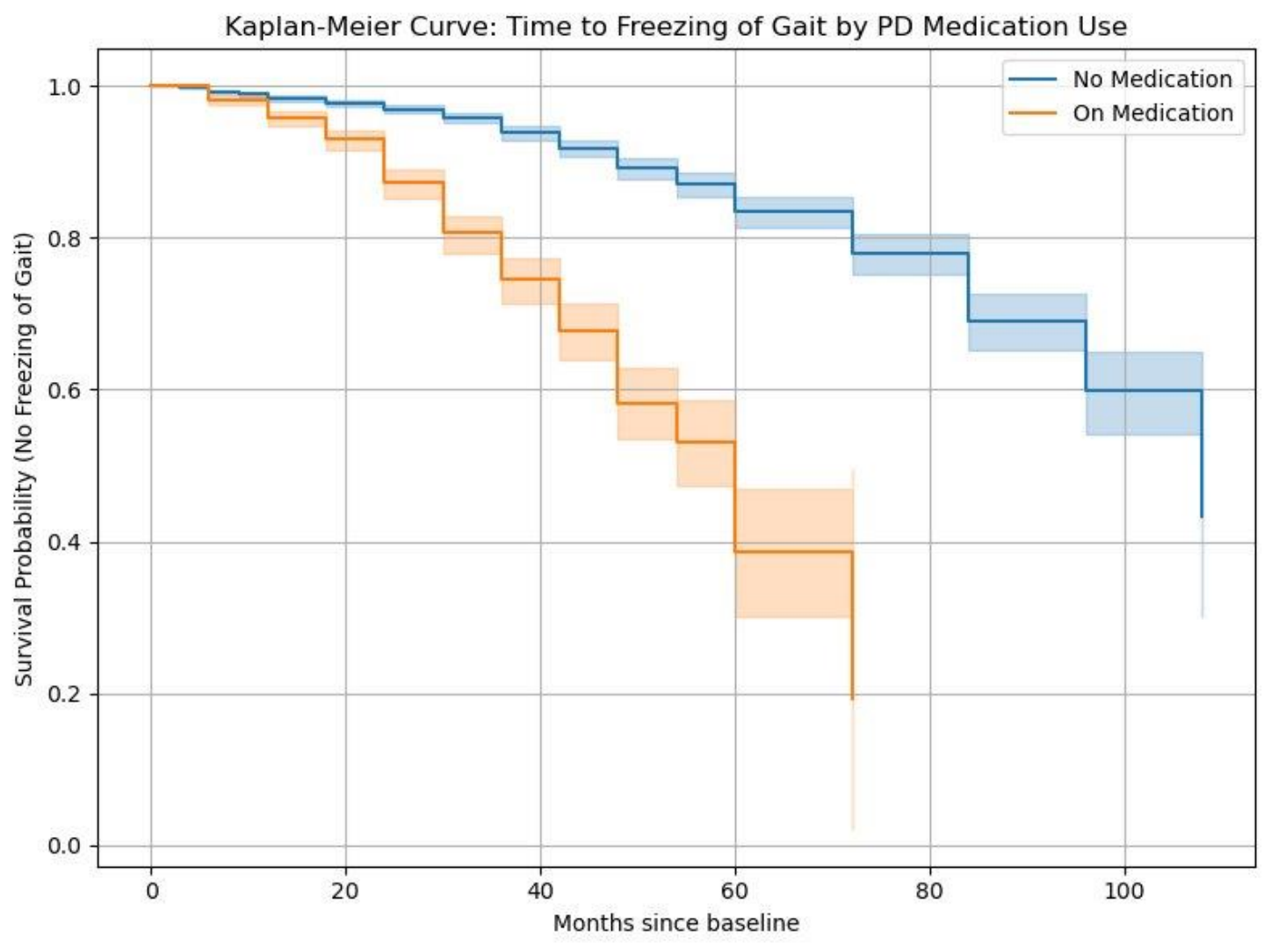
Definition: UPDRS Part 2 Q13 > 0 or UPDRS Part 3 Q11 > 0

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.



Further Directions

Our future work will focus on integrating these questions — prediction, clustering, and survival analysis — to uncover deeper relationships and shared drivers of Parkinson's disease progression

Thank You!!!