Predicting the Progression of Parkinson's Disease using Protein/Peptide Abundance Data

Shruti Raghavan
UT Austin
shrutiraghavan@utexas.edu

1 Introduction

Parkinson's disease (PD) is a chronic neurodegenerative disorder that is estimated to affect 1.2 million Americans by 2030. Patients with Parkinson's (PWP) face mobility challenges and speech and writing difficulties which lead to a negative impact on health-related quality of life and restricted independence. Research has shown that early detection of PD may help in slowing disease progression by preserving the functioning of the neurons, reducing symptoms such as difficulty in performing voluntary movements, and improved quality-of-life, and decreasing costs associated with PD (Murman, 2012). While there is no current cure for PD, treatments such as levodopa/carbidopa are more effective when administered early on in the disease (Zhu et al., 2017; noa). The pathology of PD is characterized by the aggregation of α-synuclein and an excessive loss of dopaminergic neurons (Tysnes and Storstein, 2017). Studies have shown proteins found in the cerebrospinal fluid can have different shapes in PWP and healthy individuals and could hence be potential biomarkers of PD (Karayel et al., 2022; Winchester et al., 2022; Goldman et al., 2018). The complete set of proteins involved in PD remains an open research question and any proteins that have predictive value are likely worth investigating further (Cova and Priori, 2018; Lotankar et al., 2017). In this study, we attempt to predict the progression of Parkinson's disease using protein/peptide abundance data.

The severity of PD is measured by the Unified Parkinson's Disease Rating Scale (UPDRS), which evaluates various aspects of Parkinson's disease including non-motor and motor experiences of daily living and motor complications and can be used in a clinical setting as well as in research. Our primary research question is to characterize the change in UPDRS over time by studying the shape of the trajectory of UPDRS over time (time effect). Then, we investigate whether the initial concentration of a protein/peptide influences the UPDRS trajectory (group effect). Finally, we model the time-varying relationship between UPDRS and initial protein concentration. We perform the above analyses both qualitatively and quantitatively.

Observing the evolution of UPDRS over time could help in determining whether there is an observable variation between PWP and healthy controls, which in turn would indicate whether UPDRS is an effective metric in distinguishing between the two groups. Testing for group effect enables us to evaluate the potential predictive value of a given protein/peptide as a biomarker for the early detection of PD. We could possibly identify and rank a comprehensive subset of proteins/peptides as established biomarkers for PD. Lastly, the time-varying relationship between UPDRS and initial protein/peptide concentration allows us to predict future UPDRS for new PWP and hence enables us to predict the progression of PD for new PWP.

In this study, we focus on the relationship between UPDRS_1 and the Neural cell adhesion molecule L1-like protein, hereafter represented by the code O00533. Neural cell adhesion molecules of the immunoglobulin superfamily are important components of the network of guidance cues and receptors that govern axon growth and guidance during development (Kenwrick and Doherty, 1998). An analysis of the functional capabilities of different mutated human L1-like proteins finds that it could potentially cause symptoms such as mental retardation, hydrocephalus, macrocephaly, the agenesis of the corpus callosum, spastic paraparesis, aphasia, and adducted thumbs (Hortsch, 1996). Recent research has indicated the potential role of a neural cell adhesion molecule L1-like protein as a diagnostic and prognostic marker in gastrointestinal stroma tumors (Karstens et al., 2020). This motivates us to inspect this specific molecule for potential prognostic value for PD.

2 Methods

The dataset used in this study is from the AMP®-Parkinson's Disease Progression Prediction contest on Kaggle. The dataset consists of protein abundance values derived from mass spectrometry readings of cerebrospinal fluid (CSF) samples gathered from patients over the course of multiple years while they also took assessments of PD severity. The dataset consists of three tables which are as follows:

- train_peptides.csv Mass spectrometry data at the peptide level. Peptides are the component subunits of proteins.
  - visit_id - ID code for the visit.
  - visit_month - The month of the visit, relative to the first visit by the patient.
  - patient_id - An ID code for the patient.
  - UniProt - The UniProt ID code for the associated protein. There are often several peptides per protein.
  - Peptide - The sequence of amino acids included in the peptide.
  - PeptideAbundance - The frequency of the amino acid in the sample.
- train_proteins.csv Protein expression frequencies aggregated from the peptide level data.
  - visit_id - ID code for the visit.
  - visit_month - The month of the visit, relative to the first visit by the patient.
  - patient_id - An ID code for the patient.
  - UniProt - The UniProt ID code for the associated protein. There are often several peptides per protein. The test set may include proteins not found in the train set.
  - NPX - Normalized protein expression. The frequency of the protein's occurrence in the sample. May not have a 1:1 relationship with the component peptides as some proteins contain repeated copies of a given peptide.
- train_clinical_data.csv
  - visit_id - ID code for the visit.
  - visit_month - The month of the visit, relative to the first visit by the patient.
  - patient_id - An ID code for the patient.
  - updrs_[1-4] - The patient's score for part N of the Unified Parkinson's Disease Rating Scale. Higher numbers indicate more severe symptoms. Each sub-section covers a distinct category of symptoms, such as mood and behavior for Part 1 and motor functions for Part 3.
  - upd23b_clinical_state_on_medication - Whether or not the patient was taking medication such as Levodopa during the UPDRS assessment. Expected to mainly affect the scores for Part 3 (motor function). These medications wear off fairly quickly (on the order of one day) so it's common for patients to take the motor function exam twice in a single month, both with and without medication.

Throughout this study, we consider our data to be longitudinal in nature. We investigate UPDRS1 for time, group, and time-group effects with respect to O00533.

Y_{ij} is the response variable for the ith individual $(i = 1, ..., N)$ measured at time $t_j, (j = 1, ..., n)$.

We first analyze the time effect, group effect, and interaction between time and group qualitatively by using visualization techniques suitable for longitudinal data. In order to induce groups on the different subjects, we calculate the arithmetic mean of protein concentrations at time 0 and assign 0(1) if the subject's protein concentration is greater(lesser) than the mean. Our variables for this analysis are:

- patient_id - unique id
- class_protein - a factor with categories '0' and '1'
- updrs1_t0 - UPDRS1 at 0 months
- updrs1_t12 - UPDRS1 at 12 months
- updrs1_t24 - UPDRS1 at 24 months
- updrs1_t36 - UPDRS1 at 36 months
- updrs1_t48 - UPDRS1 at 48 months

We first summarize the measures at each visit, then we visualize these summaries by the group. Next, we visualize the correlations between time points first overall and then by group.

In the second part of this study, we quantitatively analyze the time-dependent relationship between

UPDRS and initial protein concentration using Linear Mixed Models.

To use our model for inference, we make the following three general assumptions about our data:

- Assume the model: $E(Y_i) = X_i\beta$ where X_i is a vector of (time-independent) covariates
- Assume Y_i arises from a multivariate normal distribution with $Cov(Y_i) = \Sigma_i = \Sigma_i(\theta)$ where θ is a vector of covariance parameters

β is estimated using Restricted Maximum Likelihood Estimation (REML) since there are not a large number of observations per person. (The t-tests use Satterthwaite's method.) We use a Linear Mixed Model that models the covariance using a random effects covariance structure.

First, we use a random intercepts model to study the time effect as follows:

$$Y_{ij} = \beta_0+\beta_1 X_{ij_1}+\beta_2 X_{ij_2}+\beta_3 X_{ij_3}+\beta_4 X_{ij_4}+b_i+\epsilon_{ij} \tag{1}$$

where X_{ij_k} denotes measurement at month k (month 0 is used as a reference), and b_i is the random subject effect, and ϵ_{ij} is the residual error where we assume that $b \sim N(0, \sigma_b^2)$, and $\epsilon_{ij} \sim N(0, \sigma^2)$, and $b_i \perp \epsilon_{ij}$.

To model the group effect, we add a term that denotes group to the above equation:

$$\begin{aligned} Y_{ij} = \beta_0 + \beta_1 X_{ij_1}+\beta_2 X_{ij_2} + \beta_3 X_{ij_3}+ \\ \beta_4 X_{ij_4} + \beta_5 X_{ij_5}+b_i + \epsilon_{ij} \end{aligned} \tag{2}$$

where X_{ij_5} is 1 if the subject belongs to group1 (the group with O00533 level below the mean is used as a reference).

To investigate the interaction between time and group, we add interaction terms to the above model:

$$\begin{aligned} Y_{ij} = \beta_0 + \beta_1 X_{ij_1} + \beta_2 X_{ij_2} + \beta_3 \\ X_{ij_3} + \beta_4 X_{ij_4}+\beta_5 X_{ij_5} + \beta_6 X_{ij_1} X_{ij_5}+ \\ \beta_6 X_{ij_2} X_{ij_5} + \beta_7 X_{ij_3} X_{ij_5}+\beta_8 X_{ij_4} X_{ij_5} + b_i + \epsilon_{ij} \end{aligned} \tag{3}$$

We now build a linear time model by assuming a linear relationship between the time variable and the mean response allowing the linear trend varies by group:

$$Y_{ij} = \beta_0+\beta_1 X_{ij_1}+\beta_2 X_{ij_2}+\beta_3 X_{ij_1} X_{ij_2}+b_i+\epsilon_{ij} \tag{4}$$

where X_{ij_1} indicates time and X_{ij_2} indicates group and $X_{ij_1} X_{ij_2}$ is an interaction term.

Then, we visualize the subject-specific trajectories for 10 randomly chosen patients.

Finally, we attempt to fit a random intercepts and slopes model to the given data where the fixed intercept is explicit:

$$Y_{ij} = \beta_0 + X_{ij}\beta + b_i + \alpha_i X_{ij} + \epsilon_{ij} \tag{5}$$

where b_i is the random intercept, α_i is the random slope, ϵ_{ij} is the residual error and it is assumed that $\epsilon_{ij} \sim N(0, \sigma^2)$, $\alpha_i, \beta_i \perp \epsilon_{ij}$, and $(b_i, \alpha_i) \sim N((0, 0),((\sigma_b^2, \sigma_{ab}),(\sigma_{ab}, \sigma_\alpha^2))$. This model assumes that individuals vary not only in their baseline level of response (intercept) but also in terms of their changes (slope) in the mean response over time.

3 Results

From 1, we can observe that UPDRS_1 increases with month which shows that it does have a time effect. Summarizing UPDRS_1 by group (2), we find that the group with O00533 level below the mean tends to have higher UPDRS scores with the passage of time as compared to those the group with O00533 level above the mean. The positive correlation values(~ 0.7) of consecutive UPDRS measurements(4) motivate us to use linear mixed models to quantitatively measure the time effect. Visualizing consecutive UPDRS measurements by group(**??**) indicates a minimal group effect, which again is further explored quantitatively. The subject-wise trajectories (5) corroborate the findings from 2 that the increase in UPDRS_1 with time is greater in the group with O00533 level below the mean than in the group with O00533 level above the mean.

The ANOVA test has a value of 2.2e-16 and hence there is significant evidence that a random intercepts model is required. The results of the random effects model 1 shows that there is time effect: the mean UPDRS_1 at month 0 years (baseline) is 5.3521, the difference between the mean responses at 12 and 0 months is 0.4437, that between the mean responses at 24 and 0 months is 1.2887, that between the mean responses at 36 and 0 months is 1.9859, and that between the mean responses at 48 and 0 months is 2.3592. All, except the difference between 12 and 0 months, are significant. There is a significant group effect and the mean UPDRS_1 of the group with O00533 level below the mean is higher than that of the group with O00533 level above the mean by 1.5590 (table 2). The interaction of time and group is NOT significant (0.9321)

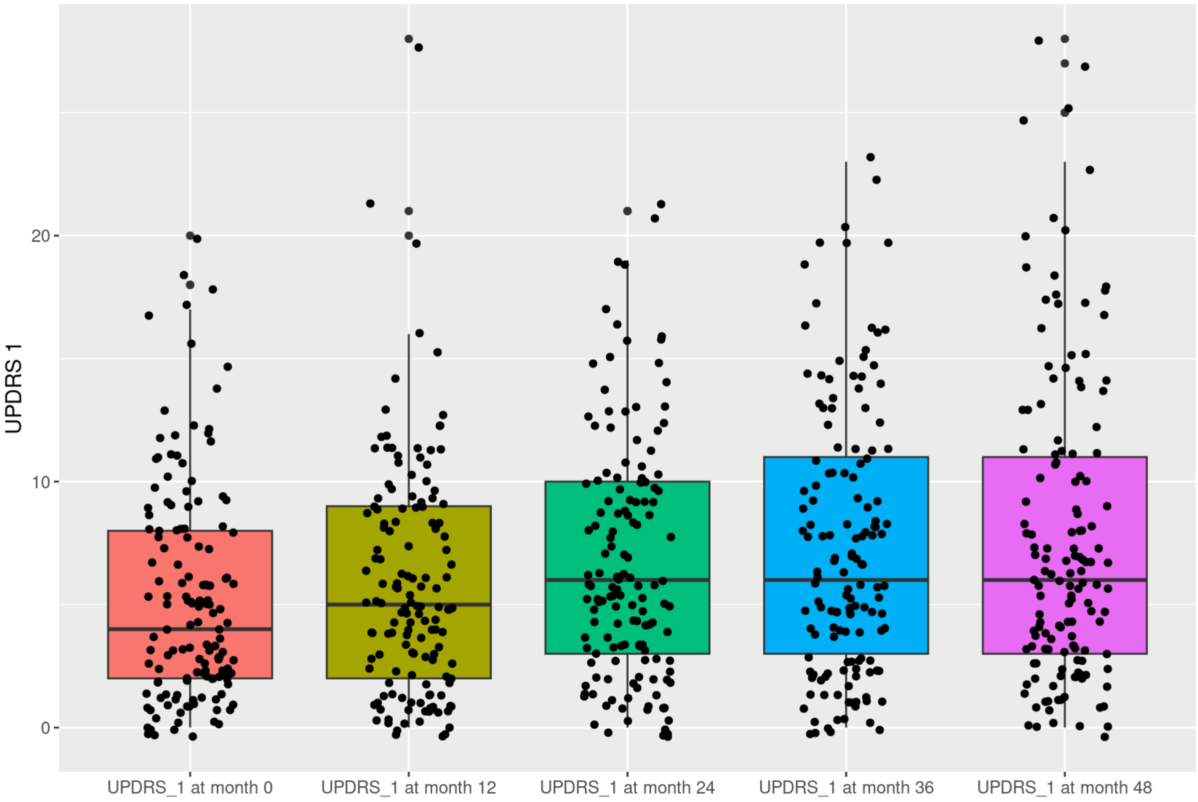

Figure 1: UPDRS_1: Summaries by Visit Month

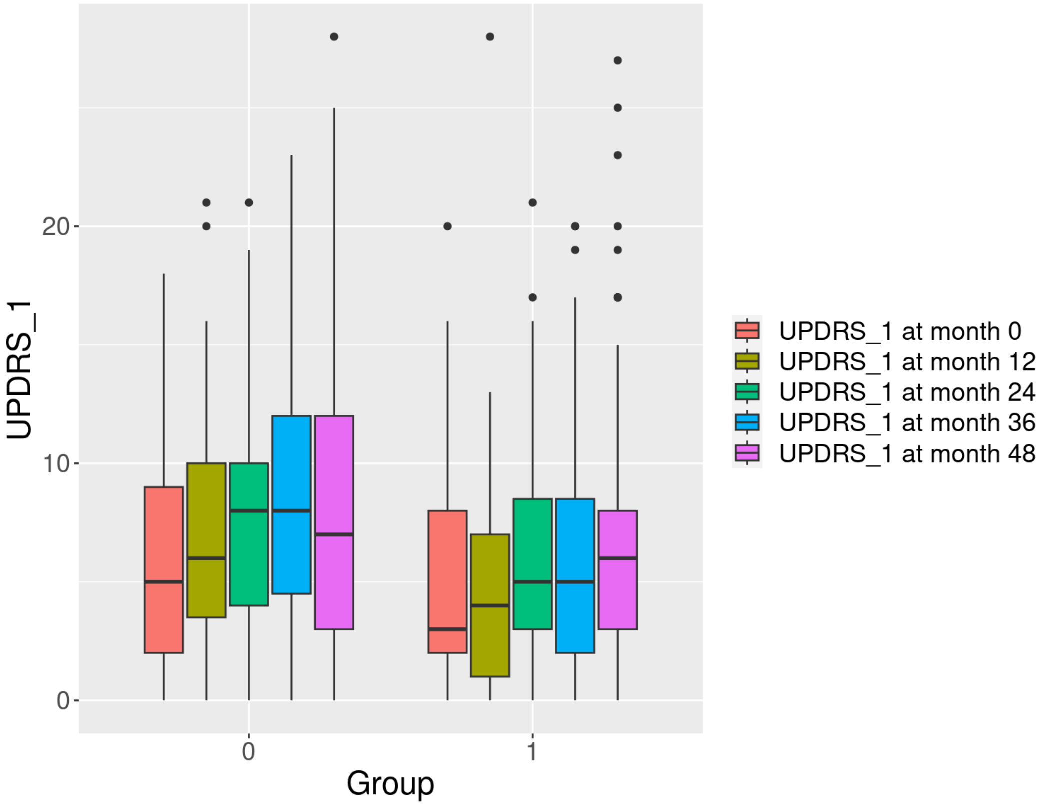

Figure 2: UPDRS_1: Summaries by Visit Month, by Group

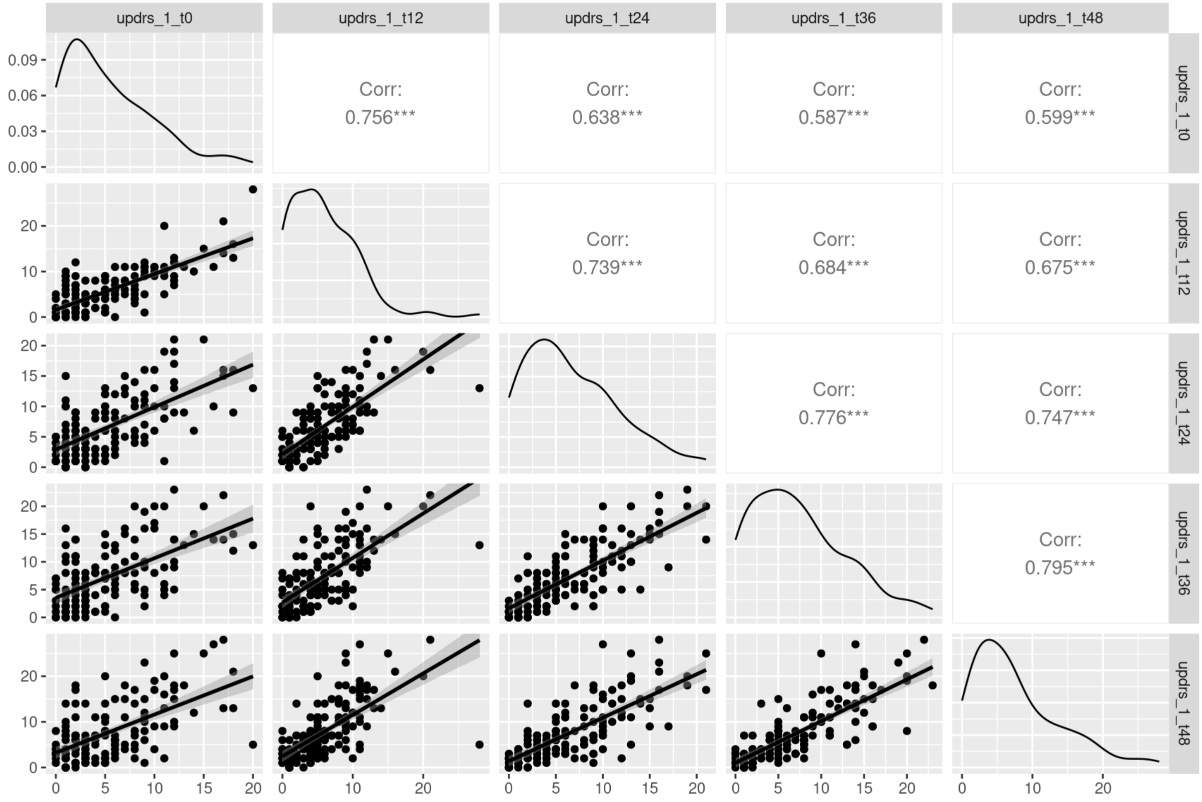

Figure 3: UPDRS_1: Visualizing Correlations over Time

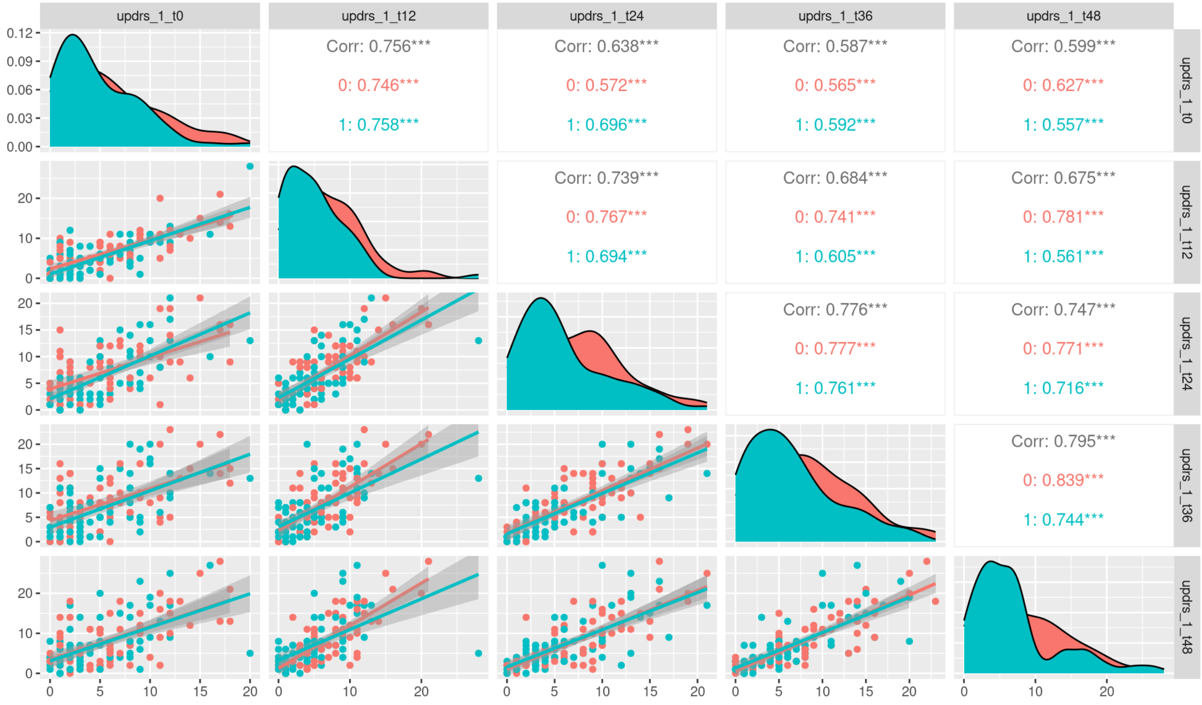

Figure 4: UPDRS_1: Visualizing Correlations over Time by Group

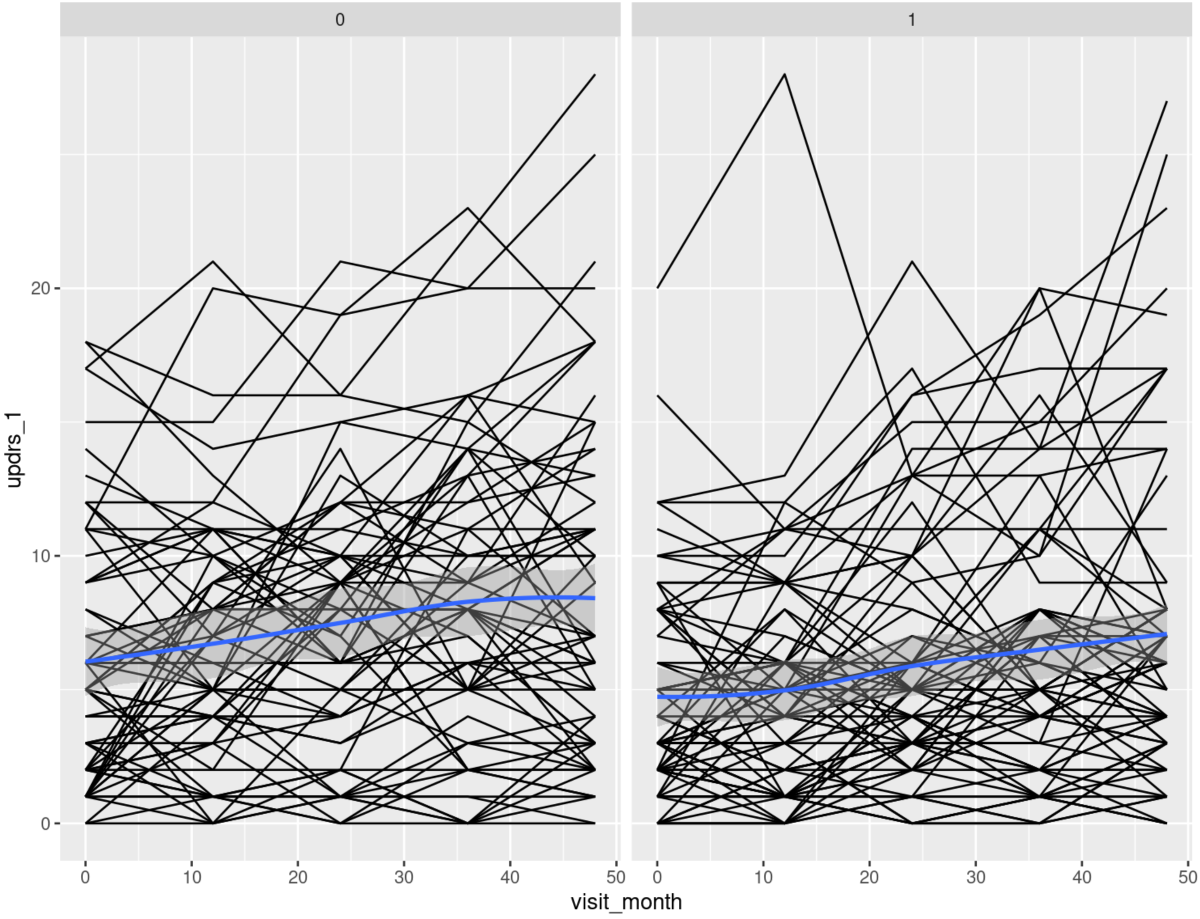

Figure 5: UPDRS_1: Subject-wise Trajectories by Group

Fixed Effects	Estimate	Std. Error	df	t value	Pr(>|t|)	Significant
Intercept	5.3521	0.4341	243.2922	12.329	<2e-16	Yes
UPDRS_1 at month 12	0.4437	0.3425	564.0000	1.295	0.195717	No
UPDRS_1 at month 24	1.2887	0.3425	564.0000	3.763	0.000186	Yes
UPDRS_1 at month 36	1.9859	0.3425	564.0000	5.798	1.12e-08	Yes
UPDRS_1 at month 48	2.3592	0.3425	564.0000	6.888	1.51e-11	Yes

Table 1: UPDRS_1: Linear Mixed Model - Time Effect

(Table 3).

The linear model 4 corroborates the findings from the linear mixed model and the initial visualizations that the group with O00533 level below the mean has higher UPDRS_1 scores than the group with O00533 level above the mean. The mean UPDRS_1 at time zero among the group with O00533 level below the mean is $\beta_0 = 6.128e + 00$ while for the group with O00533 level above the mean it is $\beta_1 = -1.539e + 00$ (the inference is that it is higher for the group with O00533 level below the mean than for the group with O00533 level above the mean). It shows that the time effect is significant (2.31e-08), whereas the group effect is not (0.9480). The subject-specific trajectories of 10 random subjects predicted by the linear time model are visualized in Figure 6.

The random slopes and intercepts failed to fit to converge to the data, perhaps because there was insufficient data to estimate both the intercepts and the parameters. Consequently, we were unable to perform any further analysis.

4 Discussion

In this study, we used visualization techniques, linear mixed models, and linear time models to investigate the time evolution of UPDRS_1 and its relationship with the concentration of the Neural cell adhesion molecule L1-like protein (coded as O00533). There was found to be a time effect and a group effect but no significant interaction between time and group. The UPDRS scores were, in general, found to increase with time, indicating an increase in the severity of PD. According to our model for group effects, the increase in UPDRS over time was higher in the group with O00533 level below the mean than in the group with O00533 level above the mean. This could possibly mean a lower concentration of the Neural cell adhesion molecule L1-like protein may indicate an increased risk for PD and that the protein is a potential biomarker for PD. Since our model uses protein concentration at time zero, it could potentially be applied for the early detection of PD.

In this study, we have analyzed the time effect, group effect, and interaction between time and group of UPDRS_1 with respect to the Neural cell adhesion molecule L1-like protein alone. This approach could easily be extended to studying all such proteins for the other three UPDRS as well. Our models incorporate protein abundance data at time zero alone, whereas there is data available for all times which would be better utilized by models that use time-varying to model time/group effects. Our random slopes and intercepts model failed to converge mostly likely due to the paucity of data. Future studies with access to more data points may be able to successfully explore such models.

Fixed Effects	Estimate	Std. Error	df	t value	Pr(>|t|)	Significant
Intercept	6.1756	0.5830	187.2441	10.593	<2e-16	Yes
UPDRS_1 at month 12	0.4437	0.3425	564.0000	1.295	0.195717	No
UPDRS_1 at month 24	1.2887	0.3425	564.0000	3.763	0.000186	Yes
UPDRS_1 at month 36	1.9859	0.3425	564.0000	5.798	1.12e-08	Yes
UPDRS_1 at month 48	2.3592	0.3425	564.0000	6.888	1.51e-11	Yes
Group1	-1.5590	0.7448	140.0000	-2.093	0.038122	Yes

Table 2: UPDRS_1: Linear Mixed Model - Group Effect

Fixed Effects	Estimate	Std. Error	df	t value	Pr(>|t|)	Significant
Intercept	6.04478	0.62687	244.76251	9.643	<2e-16	Yes
UPDRS_1 at month 12	0.67164	0.50001	560.00000	1.343	0.17973	No
UPDRS_1 at month 24	1.44776	0.50001	560.00000	2.895	0.00393	Yes
UPDRS_1 at month 36	2.23881	0.50001	560.00000	4.478	9.15e-06	Yes
UPDRS_1 at month 48	2.37313	0.50001	560.00000	4.746	2.64e-06	Yes
Group1	-1.31144	0.86256	244.76251	-1.520	0.12970	No
UPDRS_1 at month 12: Group1	-0.43164	0.68800	560.00000	-0.627	0.53066	No
UPDRS_1 at month 24: Group1	-0.30109	0.68800	560.00000	-0.438	0.66182	No
UPDRS_1 at month 36: Group1	-0.47881	0.68800	560.00000	-0.696	0.48676	No
UPDRS_1 at month 48: Group1	-0.02647	0.68800	560.00000	-0.038	0.96933	No

Table 3: UPDRS_1: Linear Mixed Model - Interaction between Time and Group

Fixed Effects	Estimate	Std. Error	df	t value	Pr(>|t|)	Significant
Intercept	6.128e+00	5.853e-01	1.901e+02	10.470	<2e-16	Yes
Group0	-1.539e+00	8.054e-01	1.901e+02	-1.911	0.0575	No
Visit Month	5.261e-02	9.283e-03	5.660e+02	5.668	2.31e-08	Yes
Group1	-8.342e-04	1.277e-02	5.660e+02	-0.065	0.9480	No

Table 4: UPDRS_1: Linear Model - Time Effect

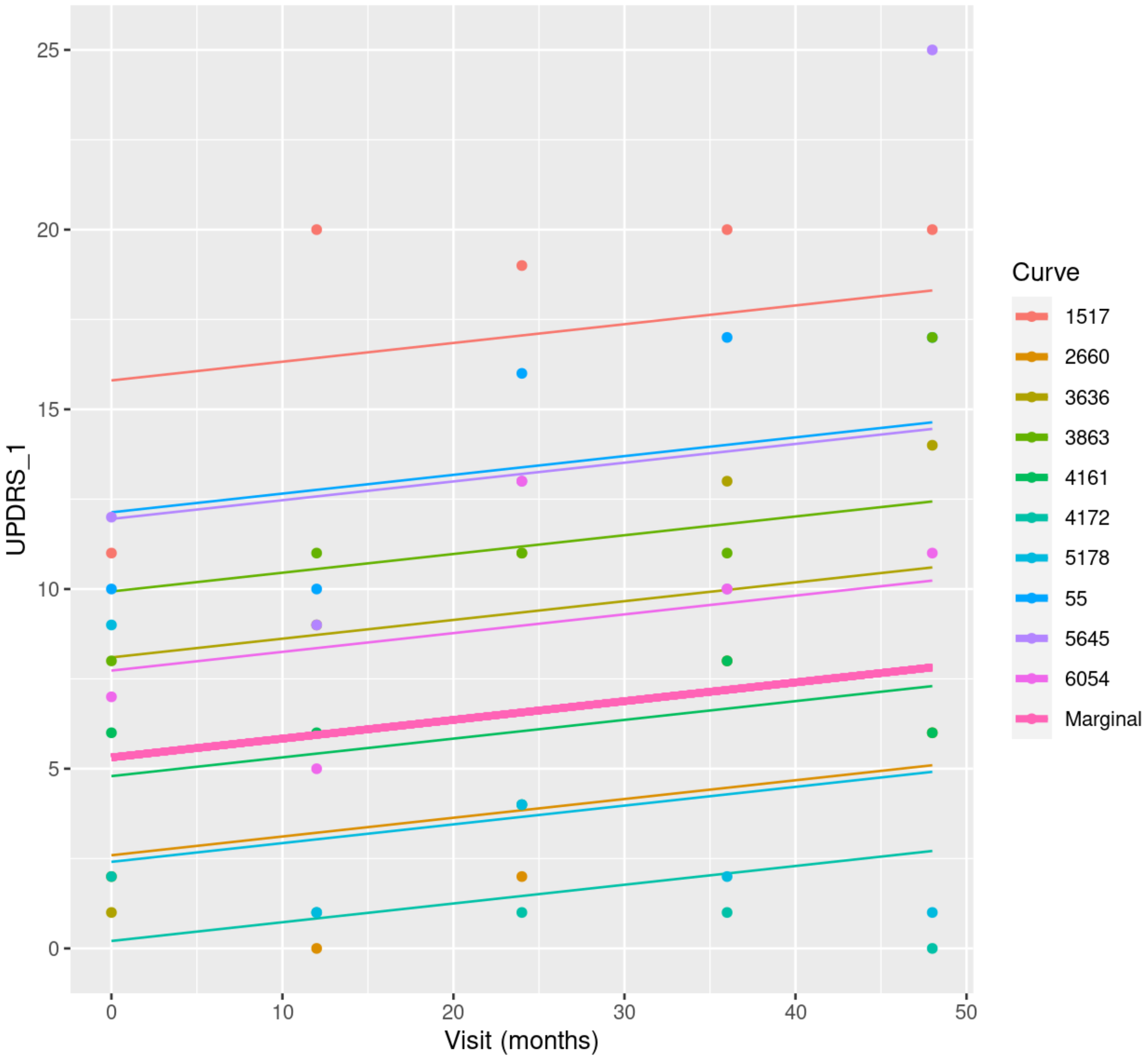

Figure 6: UPDRS_1: Subject-wise Trajectories predicted by Linear Time model

References

Levodopa and the Progression of Parkinson's Disease | NEJM.

Ilaria Cova and Alberto Priori. 2018. Diagnostic biomarkers for Parkinson's disease at a glance: where are we? *Journal of Neural Transmission*, 125(10):1417–1432.

Jennifer G. Goldman, Howard Andrews, Amy Amara, Anna Naito, Roy N. Alcalay, Leslie M. Shaw, Peggy Taylor, Tao Xie, Paul Tuite, Claire Henchcliffe, Penelope Hogarth, Samuel Frank, Marie-Helene Saint-Hilaire, Mark Frasier, Vanessa Arnedo, Alyssa N. Reimer, Margaret Sutherland, Christine Swanson-Fischer, Katrina Gwinn, The Fox Investigation of New Biomarker Discovery, and Un Jung Kang. 2018. Cerebrospinal fluid, plasma, and saliva in the BioFIND study: Relationships among biomarkers and Parkinson's disease Features. *Movement Disorders*, 33(2):282–288. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mds.27232.

Michael Hortsch. 1996. The L1 Family of Neural Cell Adhesion Molecules: Old Proteins Performing New Tricks. *Neuron*, 17(4):587–593. Publisher: Elsevier.

Ozge Karayel, Sebastian Virreira Winter, Shalini Padmanabhan, Yuliya I. Kuras, Duc Tung Vu, Idil Tuncali, Kalpana Merchant, Anne-Marie Wills, Clemens R. Scherzer, and Matthias Mann. 2022. Proteome profiling of cerebrospinal fluid reveals biomarker candidates for Parkinson's disease. *Cell Reports Medicine*, 3(6):100661.

Karl-Frederick Karstens, Eugen Bellon, Adam Polonski, Gerrit Wolters-Eisfeld, Nathaniel Melling, Matthias Reeh, Jakob R. Izbicki, and Michael Tachezy. 2020. Expression and serum levels of the neural cell adhesion molecule L1-like protein (CHL1) in gastrointestinal stroma tumors (GIST) and its prognostic power. *Oncotarget*, 11(13):1131–1140. Publisher: Impact Journals.

Sue Kenwrick and Patrick Doherty. 1998. Neural cell adhesion molecule L1: relating disease to function. *BioEssays*, 20(8):668–675. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291521-1878%28199808%2920%3A8%3C668%3A%3AAID-BIES10%3E3.0.CO%3B2-X.

Sharvari Lotankar, Kedar S Prabhavalkar, and Lokesh K Bhatt. 2017. Biomarkers for Parkinson's Disease: Recent Advancement. *Neuroscience Bulletin*, 33(5):585–597.

Daniel L. Murman. 2012. Early treatment of Parkinson's disease: opportunities for managed care. *The American Journal of Managed Care*, 18(7 Suppl):S183–188.

Ole-Bjørn Tysnes and Anette Storstein. 2017. Epidemiology of Parkinson's disease. *Journal of Neural Transmission (Vienna, Austria: 1996)*, 124(8):901–905.

Laura Winchester, Imelda Barber, Michael Lawton, Jessica Ash, Benjamine Liu, Samuel Evetts, Lucinda Hopkins-Jones, Suppalak Lewis, Catherine Bresner, Ana Belen Malpartida, Nigel Williams, Steve Gentlemen, Richard Wade-Martins, Brent Ryan, Alejo Holgado-Nevado, Michele Hu, Yoav Ben-Shlomo, Donald Grosset, and Simon Lovestone. 2022. Identification of a possible proteomic biomarker in Parkinson's disease: discovery and replication in blood, brain and cerebrospinal fluid. *Brain Communications*, 5(1):fcac343.

Huabin Zhu, Henrique Lemos, Brinda Bhatt, Bianca N. Islam, Abhijit Singh, Ashish Gurav, Lei Huang, Darren D. Browning, Andrew Mellor, Sadanand Fulzele, and Nagendra Singh. 2017. Carbidopa, a drug in use for management of Parkinson disease inhibits T cell activation and autoimmunity. *PLoS ONE*, 12(9):e0183484.