Modeling Progression of Parkinson's Disease

Liyang Sun LSUN20@MIT.EDU *MIT Economics*
Suchan Vivatsethachai SUCHANV@MIT.EDU *MIT EECS*
Christina Ji CJI@MIT.EDU *MIT EECS*

Abstract

Motor impairment is common for patients with Parkinson's disease. However, existing assessments of motor impairment are high-dimensional, making modeling difficult. In this project, we learn a low-dimensional representation that reflects the progression of motor impairment using data from Parkinson's Progression Markers Initiative (PPMI). We develop autoencoders that can accomodate the longitudinal and ordinal nature of the data. We demonstrate the latent factors identified by these autoencoders capture sufficient information to summarize patient state, are clinically interpretable, and can be used to predict progression. Our work can help doctors and patients assess how the disease will impact patient lives.

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative brain disorder [Poewe et al. 2017]. It has widespread effects, including motor, cognitive, psychiatric, and autonomic symptoms. We will focus on the hallmark motor symptoms, such as tremor, rigidity, and freezing of gait. The Movement Disorder Society developed the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) to measure these symptoms and how they affect daily activities [Rating Scales for Parkinson's Disease 2003]. However, these measurements are high-dimensional (46 questions) and ordinal (0-4 Likert scale), making modeling difficult. Existing analyses such as [Latourelle et al. 2017] simply sum the raw MDS-UPDRS scores, but this does not account for noise in each question or correlation between question responses. Instead, we learn a latent representation of the questions using autoencoders. Our study uses data from the landmark observational study Parkinson's Progression Markers Initiative (PPMI) [Marek et al. 2011], which tracks patients over several years. Our contributions are as follows:

Technical Significance Predicting progression of a large collection of symptoms is a challenging problem. It requires modeling how symptoms are related and when patients transition to a more severe state. In this project, we capture this information using a low-dimensional representation of PD that reflects severity. Specifically, we modify variational autoencoders and ordinal models to accommodate the longitudinal and ordinal nature of clinical assessments.

Clinical Relevance The low-dimensional representation helps us predict future PD conditions. Forecasting disease state and question responses helps patients assess how PD will impact their lives. Understanding whether patients will develop certain symptoms earlier or later can help with identifying groups of similar patients who may benefit from therapies or clinical trials.

2. Related work

2.1. Variational autoencoders

An autoencoder is an artificial neural network widely used to perform dimensionality reduction. The model encodes the input features into low-dimensional latent factors and then reconstructs the input from these latent factors. A commonly used alternative is the variational autoencoder (VAE) [Kingma and Welling 2014]; [Doersch 2016]. To adapt VAEs for modeling human aging, [Pierson et al. 2018] introduce monotonicity constraints on the decoder function since many biological functions tend to deteriorate monotonically over time. Since PD symptoms do not improve over time, we apply a similar monotonicity constraint in our modeling and call this the Aging model.

2.2. Ordinal regression and longitudinal loss function

Because question responses are discrete, treating them as continuous features may result in inaccurate predictions. The alternative–treating discrete features as categories in a multi-label classification problem–does not recognize some categories are closer than others. Ordinal regression models, which estimate thresholds to predict ordinal outcomes, capture this ordering. [Rennie and Srebro 2005] demonstrate that penalizing all erroneous thresholds is better than just penalizing the nearest incorrect threshold. Thus, we use their all-threshold one-sided mean squared error in our objective.

Our evaluation on two types of metrics (prediction error and ranking of latent factor by time) closely parallels the paradigm in survival analysis, where the two aims are predicting time to observed events and ordering censored events. Thus, we borrow a loss function that balances these two objectives from [Liu et al. 2018]. Details will be shown in Section 4.3.

3. Cohort and feature selection

We use only the *de novo* Parkinson's disease cohort in PPMI for two reasons: 1) These patients are all within 2 years of diagnosis and untreated at enrollment, so patient alignment and treatment confounding are not concerns. 2) The *de novo* cohort is assessed most completely and frequently. We refer readers to the PPMI study manual for eligibility criteria and assessment schedules.

We focus our analyses on untreated assessments of motor symptoms using MDS-UPDRS. Part II is a self-assessment of how PD inhibits daily activities, such as walking, speaking, and hobbies. Part III is a clinician assessment of tremor, rigidity, and motor tasks. All 46 questions are on a 0 to 4 Likert scale. For baseline features, we include demographics, biomarkers, and assessments of other symptoms. A full description of baseline features and cohort-level statistics are in Appendix A.

4. Methods

As depicted in Figure 1, we first train autoencoders to learn latent factors of input features. Then, we estimate progression rate and extrapolate to future time points. Table 1 outlines the 7 types of autoencoders we use. The baseline linear factor analysis is implemented using the factor_analyzer package [(Education Testing Services) 2019], where the number of latent

factors is selected to be the number of eigenvalues greater than 2. The remaining methods are detailed below.

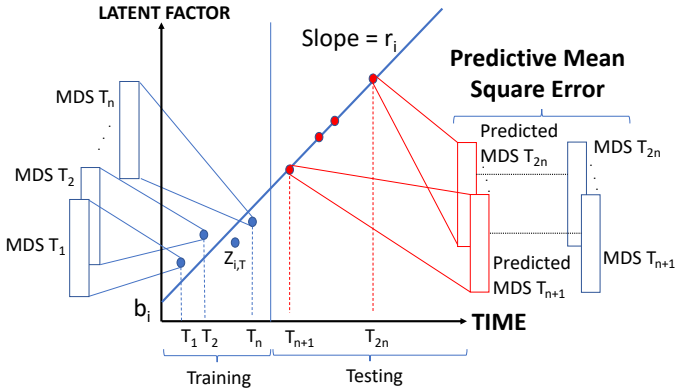

Figure 1: **Workflow overview**

Method/Property	Encoder	Decoder	Train w/ LON
Linear FA	Linear	Linear	No
VAE	Nonlinear	Nonlinear	No
Aging CS	Nonlinear	Monotonic polynomial	No
Aging LON	Nonlinear	Monotonic polynomial	Yes
Linear ordinal CS	Linear	Ordinal	No
Nonlinear ordinal CS	Nonlinear	Ordinal	No
Nonlinear ordinal LON	Nonlinear	Ordinal	Yes

Table 1: **Overview of methods**. FA: factor analysis. CS: cross-sectional. LON: longitudinal. Train w/ LON refers to training with a longitudinal loss function.

4.1. Notation

We denote input observed features as $x_{d,i,t}$ and outcomes as $y_{d,i,t}$, where d indexes the questions, i the patients, and t the visit number. The latent factors are written as $z_{i,t}$. In the ordinal model, we denote the discrete responses by k, which ranges from 0 to K, and the thresholds by θ_{dk}.

4.2. VAE and Aging Model

The dataset used in [Pierson et al. 2018] has only one timepoint per patient. To utilize the longitudinal data in PPMI, we consider two approaches: 1) We treat data points from the same patient as independent and apply the original Aging model. 2) We modify the original Aging model to incorporate longitudinal data while learning. We refer to our modification as the Aging longitudinal model (Aging LON) and their original as the Aging cross-sectional model (Aging CS). We also use a regular VAE as a nonlinear baseline. VAE-based models require large amounts of data, but we have less than 2000 points in our training set. Thus, we synthesize intermediate timepoints via linear interpolation of adjacent timepoints. The details are in Appendix B.

4.3. Ordinal regression

We use the all-threshold one-sided square loss from [Rennie and Srebro 2005]. We modify their model to set thresholds for multiple ordinal questions using the same latent factor. Additionally, we allow for nonlinear encoders and incorporate the longitudinal loss from [Liu et al. 2018]. The combined objective function is

$$\begin{aligned} l(x,y) = &\sum_{d,i,t}\sum_{k=1}^{K}\left(\left(z_{i,t}-\theta_{d,i,t}\right)^2\left(\mathbb{1}\left\{y_{d,i,t}\geq k, x_{d,i,t}<k\right\}+\mathbb{1}\left\{y_{d,i,t}<k, x_{d,i,t}\geq k\right\}\right)\right) \\ &-\frac{\alpha}{\sum_i\left|\mathcal{E}^{(i)}\right|}\left(\sum_i\sum_{(p,q)\in\mathcal{E}^{(i)}}\log\sigma\left(z_{i,p}-z_{i,q}\right)\right), \end{aligned} \tag{1}$$

where σ is the sigmoid function, α is a non-negative constant, and $\mathcal{E}^{(i)}$ is the set of all pairs of ordered timepoints, i.e. existence of an edge in $\mathcal{E}^{(i)}$ implies time index $p > q$ for patient i. We vary α to tune the weight of the ranking loss. Setting $\alpha = 0$ gives us the cross-sectional model.

4.4. Evaluation metrics

First, we examine the latent factors our models discover by studying how they correlate with the observed features. Second, we consider four quantitative metrics that measure how informative the latent factors are to the PD progression.

4.4.1. Concordance index and consecutive visit ranking

We measure how well the latent factors obey time ordering using the concordance index (CI) and consecutive visit ranking. CI is given by

$$\frac{1}{\sum_i\left|\mathcal{E}^{(i)}\right|}\sum_i\sum_{(p,q)\in\mathcal{E}^{(i)}}\mathbb{1}\left\{z_{i,p}>z_{i,q}\right\}, \tag{2}$$

Existence of an edge in $\mathcal{E}^{(i)}$ implies time index $p > q$ for patient i. Consecutive visit ranking is a more stringent metric since the average is taken over the set of adjacent visits $\mathcal{E}^{(i)}_{t,t+1}$ instead. For models with multiple latent factors, we calculate CI for each latent factor, but only report the highest here, assuming that other latent factors may be capturing non-temporal information.

4.4.2. Reconstruction and prediction mean squared error

We evaluate how well the autoencoders can reconstruct input features. Specifically, we calculate the mean squared error (MSE) between the reconstructed and observed features. For prediction, we first fit a patient-specific linear regression of latent factor on time since enrollment. The first half of timepoints are used for fitting; the second half for prediction. Feeding the extrapolated latent factors into the decoder outputs predictions for the question responses. MSE can then be calculated between the predictions and observations. Note that MSEs for ordinal models are necessarily larger since ordinal models cannot output non-integer values.

5. Results

5.1. Learned latent factors and thresholds

As seen in Figure 2, the latent factors show a generally increasing trend but with noisy fluctuations. Figure 3 shows that when there are one or two latent factors, almost all the features are positively correlated with them, as all the MDS-UPDRS II and III questions are generally correlated. With multiple latent factors or hidden units, the VAE is best at disentangling as expected, but the features related to each latent factor do not form coherent subgroups.

Table 2 shows that the VAE performs better than the linear baseline on MSEs, indicating that nonlinearity is indeed necessary. The cross-sectional ordinal models perform best on the ranking metrics since a thresholding function is monotonic. The cross-sectional Aging model performs the best all-around, since it captures the benefits of nonlinearities as in the VAE and monotonicity from its polynomial decoder. The decreased performance of both longitudinal models requires further analysis.

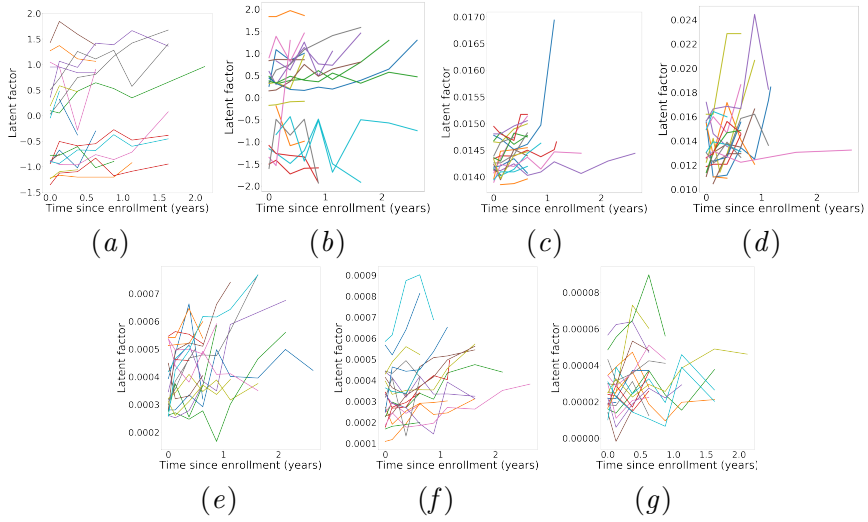

Figure 2: **Latent factors across time** for 20 randomly sampled patients in test set. a) First latent factor in linear FA. b) VAE. c) Aging CS. d) Aging LON e) Lin ord CS. f) Nonlin ord CS. g) Nonlin ord LON.

Because the latent factor is correlated with time since enrollment, we can interpret question responses with lower thresholds as symptoms that tend to occur earlier for patients. For example, as seen in Figure 4, larger rest tremor amplitude in the upper left body and rigidity in the lower right body might appear earlier than difficulty getting up from a deep chair, impairment of hobbies, or postural tremor in the right hand.

5.2. Subtyping

To separate patients into slow and fast progressors, we obtain rates by fitting patient-specific linear regressions of the first latent factor on time using their entire trajectory. Then, we divide patients at the median rate of the training set. We compare how the distributions

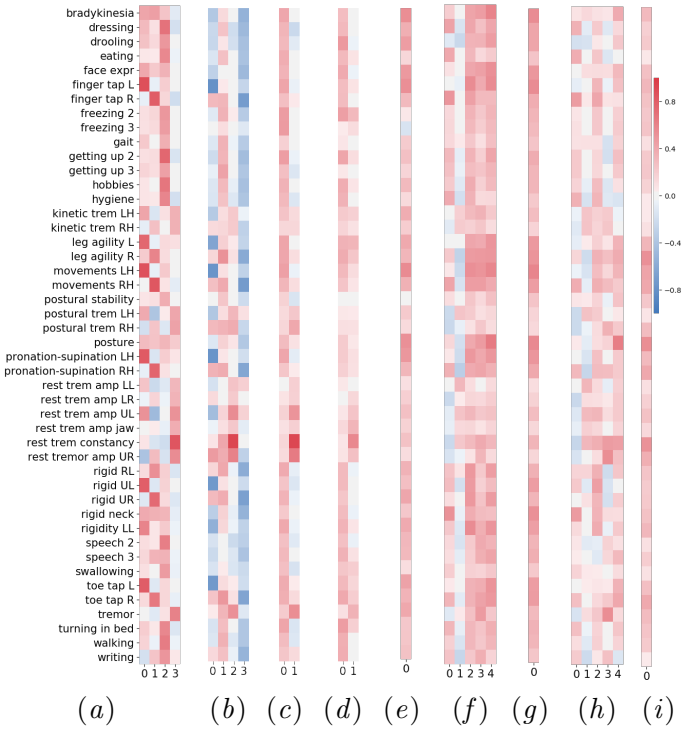

Figure 3: **Correlation between latent factors and observed features** in each model. a) Linear FA. b) VAE. c) Aging CS. d) Aging LON. e) Lin ord CS. f) Nonlin ord CS hidden. g) Nonlin ord CS latent. h) Nonlin ord LON hidden. i) Nonlin ord LON latent.

Method/Metric	CI	Consec. rank	MSE recons.	MSE pred.
Linear FA	0.682 (0.010)	0.577 (0.006)	0.862 (0.010)	1.533 (0.030)
VAE	0.570 (0.026)	0.528 (0.016)	**0.392** (0.016)	1.344 (0.538)
Aging CS	0.727 (0.036)	0.616 (0.020)	0.559 (0.016)	**0.631** (0.018)
Aging LON	0.657 (0.031)	0.561 (0.013)	0.666 (0.084)	0.846 (0.107)
Lin ord CS	0.730 (0.026)	**0.619** (0.028)	1.504 (0.015)	1.856 (0.089)
Nonlin ord CS	**0.739** (0.017)	**0.619** (0.025)	1.302 (0.013)	1.582 (0.051)
Nonlin ord LON	0.651 (0.011)	0.554 (0.011)	1.897 (0.038)	2.163 (0.078)

Table 2: **Performance of methods** on test set in 5-fold cross-validation. Mean from 5 folds is shown, followed by standard deviation in parentheses. Consec. rank: consecutive visit ranking. MSE recons.: MSE on reconstructing input data. MSE pred.: MSE on predictions.

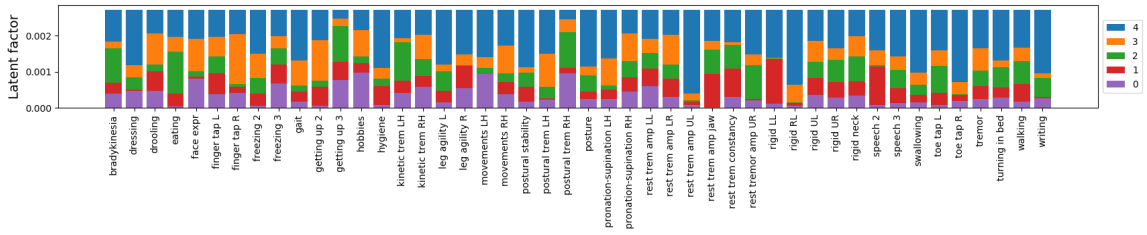

Figure 4: **Ordinal thresholds** from nonlinear cross-sectional model.

Characteristic	Cluster 0 mean	Cluster 1 mean	p-value
Latent 0 rate	-0.283	1.020	7.943e-41
Latent 1 rate	-0.035	0.753	1.896e-11
Right-dominant	0.325	0.530	5.423e-05
DaTscan ipsilateral putamen	1.017	0.900	0.003
DaTscan ipsilateral caudate	2.219	2.068	0.013
DaTscan contralateral putamen	0.734	0.659	0.007
DaTscan contralateral caudate	1.909	1.765	0.013
CSF pTau to tTau	0.080	0.073	0.012
CSF pTau log	2.224	1.664	0.013
CSF alpha-synuclein log	7.270	7.180	0.038
Epworth sleep	0.194	0.110	0.025
GDS depressed	0.110	0.188	0.036
TD	0.827	0.823	0.919
PIGD	0.042	0.083	0.104

Table 3: **Cluster profiles** at baseline from linear factor analysis model. p-value is from Welch's t-test. All features with p-value less than 0.05 are shown. Patients who are neither TD nor PIGD are indeterminate.

of baseline features differ between the two clusters and train a logistic regression predicting cluster membership from baseline features using sklearn [Pedregosa et al. 2011]. Baseline features are min-max normalized to 0-1 using the training range so coefficients are comparable.

As seen in Table 3, progression appears to be faster among patients who are affected more severely by the disease on the right side of their body. This may be because most people are right-dominant, and as such, the disease is more likely to hinder their daily life. Lower dopamine levels in DaTscan brain imaging, lower levels of some CSF biomarkers, and depression are associated with faster progression. A motor subtyping system that is well-established in PD research is tremor-dominant (TD) vs postural instability gait-dominant (PIGD) [Stebbins et al. 2013]. We find that the proportion of patients who are PIGD is higher in the faster-progressing group, aligning with the review by [Xia and Mao 2012], but the difference is not statistically significant.

Comparing Tables 3 and 4, the features that have large coefficients in the classifier differ from those with significantly different distributions across the two clusters. The negative sign on baseline MDS-UPDRS III and absence of age contradict established knowledge. Genetic PCA component 0, which captures single nucleotide polymorphisms associated with catecholamine O-methyltransferase (COMT), has a large positive coefficient. COMT is an enzyme that breaks down dopamine and other neurotransmitters and a frequent PD treatment target, so this result seems sensible. Table 6 in Appendix C shows that the classifiers perform similarly for all models. The cluster profiles and coefficients of other models are omitted for brevity.

6. Discussion

We adapt autoencoders specifically for disease progression. First, we extend the Aging model from [Pierson et al. 2018] to incorporate longitudinal data. This Aging longitudi-

Feature	Coeff	Feature	Coeff
Right-dominant	1.152	DaTscan ipsilateral putamen	-1.183
GDS depressed	0.988	DaTscan contralateral putamen	-1.066
HVLT retention	0.655	CSF pTau (log-scaled)	-1.043
SCOPA-AUT	0.640	Height (cm)	-0.998
# years education	0.571	MDS-UPDRS III	-0.948
PIGD	0.530	Epworth sleep	-0.883
Right-handed	0.420	Letter-number sequencing	-0.496
Genetic PCA comp 0	0.386	RBD sleep disorder	-0.280
Male	0.297	Symbol-digits modality	-0.275

Table 4: Top 9 largest and smallest **coefficients in logistic regression** predicting cluster membership from baseline features. Clusters are determined using the linear factor analysis model.

nal model can be used in any disease progression task where symptoms are expected to be monotonic and multiple timepoints are collected for each patient. Second, we modify the decoder function of an autoencoder to be multiple ordinal regressions. Even with a small dataset, these two models can still capture meaningful latent factors for PD motor symptoms. These factors are particularly correlated with hand movements and rest tremor constancy. They can be predicted by right-dominant PD and imaging and CSF biomarkers at baseline. These latent factors can be used to stage PD patients so that the clinician can assign appropriate treatment. However, we need further clinical validation.

6.1. Limitations

A limitation of the deterministic single latent factor ordinal model is that it cannot model noise in question responses and satisfy the constraints imposed by multiple questions. A limitation to all methods is that the data distribution, specifically the relationship between latent factors and observed features, shifts across time. Because all patients in the *de novo* cohort are within a few years of diagnosis, this is not a problem in the dataset we test on. However, this may be an issue if we apply our model to later-stage PD patients.

6.2. Future work

We can address the limitation above for ordinal models by using multiple latent factors, where each latent factor is tied to a subgroup of related questions. We can also implement a probabilistic model. More broadly, because most PD patients eventually start treatment, we should modify all our models to account for treatment. This would vastly increase the clinical utility of our model. We can also expand the set of input features to include other assessments. Finally, validating our findings on another PD dataset and applying our methods to other diseases or longitudinal survey datasets can broaden our impact.

7. Division of labor

Liyang explored reasonable baseline models and implemented the linear factor model. She set up part of the evaluation and classification scripts. She also designed and worked on making the poster.
Suchan modified VAE and Aging models to fit PPMI data. He extended Aging model to incorporate longitudinal data. He also designed and implemented a sampled function that generates more data points using linear interpolation.
Christina pre-processed the PPMI data, derived and implemented the ordinal models, set up part of the evaluation and classification scripts, and helped build a pipeline for the entire workflow. She also worked on writing the paper.

Acknowledgement

We thank Professor David Sontag for guiding us in this project. We also thank Charles Venuto and Monica Javidnia for sharing their clinical expertise. Data used in the preparation of this article were obtained from the Parkinsons Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI-a public-private partnership-is funded by the Michael J. Fox Foundation for Parkinsons Research and funding partners, including abbvie, Allergan, Avid, Biogen, BioLegend, Bristol-Myers Squibb, Denali, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pzer, Piramal, Prevail, Roche, Sano Genzyme, Servier, Takeda, Teva, UCB, and Golub Capital.

References

Rating Scales for Parkinson's Disease, Movement Disorder Society Task Force on (2003). "The unified Parkinson's disease rating scale (UPDRS): status and recommendations". In: *Movement Disorders* 18.7, pp. 738–750.

Rennie, Jason DM and Nathan Srebro (2005). "Loss functions for preference levels: Regression with discrete ordered labels". In: *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*. Kluwer Norwell, MA, pp. 180–186.

Marek, Kenneth et al. (2011). "The parkinson progression marker initiative (PPMI)". In: *Progress in neurobiology* 95.4, pp. 629–635.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Xia, Ruiping and Zhi-Hong Mao (2012). "Progression of motor symptoms in Parkinson's disease". In: *Neuroscience bulletin* 28.1, pp. 39–48.

Stebbins, Glenn T et al. (2013). "How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: comparison with the unified Parkinson's disease rating scale". In: *Movement Disorders* 28.5, pp. 668–670.

Kingma, Diederik P. and Max Welling (2014). "Auto-Encoding Variational Bayes". In: *CoRR* abs/1312.6114.

Doersch, Carl (2016). "Tutorial on variational autoencoders". In: *arXiv:1606.05908*.

Latourelle, Jeanne C et al. (2017). "Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed Parkinson's disease: a longitudinal cohort study and validation". In: *The Lancet Neurology* 16.11, pp. 908–916.

Poewe, Werner et al. (2017). "Parkinson disease". In: *Nature reviews Disease primers* 3, p. 17013.

Liu, Bin et al. (2018). "Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Pierson, Emma et al. (2018). "Inferring Multidimensional Rates of Aging from Cross-Sectional Data". In: *arXiv:1807.04709 [cs, stat]*. arXiv: 1807.04709. URL: http://arxiv.org/abs/1807.04709 (visited on 03/17/2019).

(Education Testing Services), jbiggsets (2019). *Factor Analyzer*. https://github.com/EducationalTestingService/factor_analyzer.

Appendix A.

The baseline features consist of the following 62 features:

- Demographics: Male, white, age, number of years of education, right-handed
- Basic Parkinson's features: has family history of the disease, time since diagnosis, dominant side affected by disease (right or left)
- Vitals: Systolic and diastolic blood pressures (standing and supine), heart rate (standing and supine), temperature, weight, height
- Smell: UPenn smell identification test total
- Genetic SNPs (single nucleotide polymorphisms): Using 193 SNPs that were measured for at least 90% of PD patients, we performed PCA and took the top 10 components.
- DaTscan (Dopamine active transporter scan) imaging: dopamine levels in the ipsilateral (same as dominant disease side) and contrateral (opposite) sides of the putamen and caudate brain regions, asymmetry indices between the 2 sides, and count density ratios.
- Cerebrospinal fluid (CSF) biomarkers: phosphorylated Tau (pTau, log-scaled), total Tau (tTau, log-scaled), alpha-synuclein (log-scaled), A-beta (log-scaled), pTau to tTau ratio, tTau to A-beta ratio, pTau to A-beta ratio, hemoglobin
- MDS-UPDRS: parts I, II, III totals
- Binary indicators for TD vs PIGD calculated using MDS-UPDRS questions
- Cognitive assessments: MoCA (Montreal cognitive assessment), BJLO (Benton judgment of line orientation), LNS (letter-number sequencing), semantic fluency, symbol-digit modalities test, HVLT (Hopkins verbal learning test) retention, HVLT discrimination recognition, HVLT immediate recall

Feature	Mean (SE)	Feature	Percentage
# visits	5.3 (2.4)	Male	65.5%
Disease duration	0.45 (0.53)	White	94.8%
Age	61.6 (9.7)	Family history	24.3 %
UPSIT	22.4 (8.2)	Right-handed	88.7%
MDS-UPDRS I	5.8 (4.2)	Right-dominant	42.3%
MDS-UPDRS II	5.7 (4.2)	GDS depressed	13.9%
MDS-UPDRS III	20.3 (8.9)	EPWORTH sleep	15.6%
MoCA	27.1 (2.3)	REM sleep	37.6%
SCOPA-AUT	9.5 (6.2)	QUIP	20.6%
STAI	65.2 (18.6)		

Table 5: **Cohort characteristics** for the 423 de novo Parkinson's disease patients in PPMI at baseline. # visits refers to the number of visits with untreated MDS-UPDRS exams. Disease duration is time since diagnosis in years. Right-dominant means Parkinson's affects the right side more severely. QUIP is the proportion of patients with at least 1 impulsive behavior. Not shown: 2.3% of patients are ambidextrous. 2.4% of patients are affected equally by Parkinson's on both sides of the body. The remaining are left-handed and left-dominant, respectively.

- Sleep: Epworth daytime sleepiness, REM sleep behavior disorder
- Psychiatric: STAI (state-train anxiety inventory), GDS (geriatric depression scale), QUIP (questionnaire for impulsive-compulsive behavior in PD patients
- Autonomic: SCOPA-AUT (scales for outcomes in Parkinson's disease-autonomic)

Cohort-level statistics on a subset of these features are shown in Table 5. Among the 423 PD patients, 372 have a complete set of baseline features. The remaining are missing either imaging or CSF biomarkers. For cluster classification, we use the smaller set of patients with complete baseline features.

Appendix B.

Assume that for patient i, we observe features of D questions $(x_{d,i,t} \mid d = 1, \ldots, D)$ for time $t = t_1, \ldots, t_T$. To sample a data point for time t' such that $t_l < t' < t_{l+1}$, we approximate the score $x_{d,i,t'}$ by linear interpolation from the adjacent observed score x_{d,i,t_l} and $x_{d,i,t_{l+1}}$. Specifically, we set

$$x_{d,i,t'} = x_{d,i,t_l} + (t' - t_l)\frac{x_{d,i,t_{l+1}} - x_{d,i,t_l}}{t_{l+1} - t_l}$$

for all $d = 1, \ldots, D$. An example of interpolated scores for sampled data points is illustrated in Figure 5.

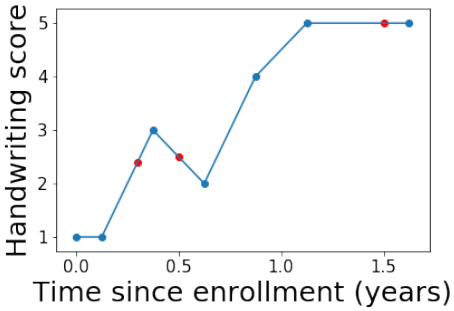

Figure 5: Handwriting scores of a patient across time. The blue dots are the scores of the real observed data point. The red dots are the interpolated scores of the sampled data points

Appendix C.

Method	AUROC	Accuracy	Precision	Recall
Linear FA	0.655 (0.067)	0.599 (0.059)	0.598 (0.063)	0.582 (0.067)
VAE	0.637 (0.067)	0.614 (0.082)	0.619 (0.104)	0.617 (0.105)
Aging CS	0.596 (0.083)	0.452 (0.138)	0.712 (0.361)	0.318 (0.242)
Aging LON	0.656 (0.053)	0.538 (0.057)	0.861 (0.054)	0.466 (0.086)
Lin ord CS	0.550 (0.054)	0.521 (0.032)	0.513 (0.081)	0.544 (0.087)
Nonlin ord CS	0.547 (0.014)	0.527 (0.045)	0.531 (0.036)	0.510 (0.041)
Nonlin ord LON	0.517 (0.031)	0.522 (0.050)	0.531 (0.093)	0.553 (0.057)

Table 6: **Evaluation of cluster prediction**. Mean from 5-fold cross-evaluation is shown, followed by standard deviation in parentheses. Precision and recall are for cluster 1.