Title: Comprehensively modeling heterogeneous symptom progression for Parkinson's disease subtyping

Authors: Chang Su, PhD[1]; Yu Hou, PhD[2]; Matthew Brendel, BSc MEng[3]; Claire Henchcliffe, MD, DPhil[4]; Fei Wang[*], PhD[2]

[1]Department of Health Service Administration and Policy, Temple University, Philadelphia, PA, USA.
[2]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA.
[3]Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, New York, NY, USA.
[4]Department of Neurology, University of California Irvine, Irvine, CA, USA
[*]Corresponding author

Corresponding author:
Fei Wang,
425 E 61 St. New York City. NY 10065. USA
few2001@med.cornell.edu

In total 4790 words

Abstract

The Parkinson's disease (PD) is a heterogeneous neurodegenerative disease, of which the etiological and pathological mechanisms remain unclear to date. PD has been associated with diverse movement dysfunctions and non-motor symptoms (i.e., symptom heterogeneity) and progression patterns of these symptoms differ from patient to patient (i.e., progression heterogeneity). To address these, the present investigation aims at comprehensively considering full progression course of early PDs to identify subtypes, each of which can reflect unique PD progression pattern. We retrospectively analyzed the Parkinson's Progression Markers Initiative (PPMI) and the Parkinson Disease Biomarkers Program (PDBP) as the development and validation cohorts, respectively. An unsupervised deep learning model was built to model progression trajectories in diverse clinical manifestations and cerebrospinal fluid (CSF) biomarkers to produce a representation vector for each patient, encoding his/her symptom progression profile. Then by performing clustering analysis on the patients' representation vectors, we identified three subtypes with distinct PD progression patterns in the PPMI cohort: Subtype I, mild baseline severity and mild symptom progression; mild baseline severity and moderate progression; and Subtype III, rapid symptom progression. Replication in the PDBP validation cohort demonstrated reproducibility of the subtypes. After that, we explored demographic factors, CSF biomarkers, neuroimaging biomarkers in brain regional atrophy, and genetic factors of the subtypes. Last, to enhance usability of the subtypes, predictive model of subtypes that relies on data at baseline and 1-year follow-up was trained. In conclusion, the identified subtypes revealed significant symptom progression patterns of PDs. Patients with similar baseline severities can even suffer from different progression pattern, leading to distinct prognosis. Demographic factors, biomarkers, and genetic components of the subtypes suggested distinct biological mechanisms and pathways potentially leading to those progression patterns. Our findings may benefit pathophysiological study, clinical practice, and clinical trials to advance PD.

Introduction

Parkinson's disease (PD) is a chronic heterogenous degenerative disorder of the human central nervous system, of which patients typically suffer from progressive motor and non-motor dysfunctions[1,2]. To date, the etiological and pathological mechanisms of PD remain unclear and there is no disease-modifying treatment in most cases[3]. In this context, the identification of subtypes of a specific disease has become a crucial approach to address complex diseases like PD.

In PD, the critical attributes that hinder researchers to better understand the disease include the symptom heterogeneity[4,5], i.e., patients are usually associated with a wide spectrum of motor (like tremors, rigidity, bradykinesia, etc.) and non-motor symptoms (like cognition dysfunction, sleep disorder, mental disorder, etc.), as well as the progression heterogeneity[5], i.e., these symptoms progress gradually and the progression patterns (e.g., annual progression rates) differ from patient to patient. Conventional PD subtyping efforts, which generally classified subtypes relying on hypothesized criteria based on specific predominant clinical features like tremor, postural instability, cognition, depression, sleep disorders, or age at onset, etc., are limited due to potentially biased to the feature used or the hypothesis predefined[6]. Recently data-driven subtyping approaches have been engaged to address these issues via retrospectively modeling patients' historical records to identify clinically meaningful subtypes[6,7]. Typically, previous data-driven efforts relied on clustering analysis (or combined with some feature engineering techniques like dimension reduction) to identify subtypes using baseline data or data collected at baseline and follow-up time points of patients[6,7]. Although these studies have resulted in meaningful subtypes that showed distinct baseline clinical characteristics and prognostic patterns, the resultant subtypes were not able to accurately describe progression patterns of PD, mainly because that the methods used were not capable to uncover underlying patterns from the multivariate clinical trajectories of patients. In this context, advanced techniques, like deep learning approaches, which were specifically designed to modeling multivariate time sequence data for hidden pattern discovery[8,9], could be the optimal way to solve this problem.

In this investigation, our goal was to identify PD progression subtypes, such that patients of a subtype share similar progression patterns across the PD symptoms. To achieve this, we used early PDs of two longitudinal cohorts, the Parkinson's Progression Markers Initiative (PPMI)[10] and the Parkinson Disease Biomarkers Program (PDBP)[11]. We extended our previous effort[12] and proposed an advanced deep learning model that can comprehensively account full course progression in diverse clinical manifestations and CSF biomarkers of early PDs and produce a representation vector capturing the progression pattern of the patients. Compared to our previous model[12], the new model was trained in an unsupervised manner, enabling it to discover progression patterns more robustly with less bias. Using the learned patient representation vectors, we performed clustering analysis and identified PD subtypes with significant progression patterns in the PPMI cohort. Replication in the independent PDBP cohort demonstrated the reproducibility of the subtypes. In addition, we conducted statistical analyses to explore a wide range of demographic factors, cerebrospinal fluid (CSF) and neuroimaging biomarkers (within 1-year follow-up), and genetic components of the subtypes, which may add significant prognostic implications. Last, to further enhance usability of the subtypes in clinical practices, predictive models that can predict patients' subtype memberships using data within 1-year follow-up were built. The overall workflow of our study is illustrated in Figure 1.

Results

Patients

We used the PPMI[10] cohort, an international and multi-center observational study that longitudinally and comprehensively collects data and bio-samples from *de novo* PDs (PD diagnosis within the last 2 years and untreated at enrollment), healthy control (HC) volunteers, and at-risk populations. We included 406 early PD patients in the PPMI cohort, consisting of 140 (34.5%) females and 266 (65.5%) males with an age of 59.6±10.0 years at PD onset and PD durations of 0.6±0.7 year at baseline (see Supplementary Table 1). Besides, 188 healthy controls (HCs) and 61 subjects with scans

without evidence for dopaminergic deficit (SWEDD) were also included for model training (see Supplementary Table 1).

We used PDBP[11] as the independent validation cohort, which longitudinally and comprehensively collects data and bio-samples from patients with PDs and HC participants. In the PDBP cohort, we included 210 early PDs (symptom duration < 3 years at enrollment) consisting of 90 (42.9%) females and 120 (57.1%) males with an age of 63.0±10.0 years at PD onset and PD durations of 1.0±0.8 year at baseline (see Supplementary Table 1). Besides, 211 HCs and 287 other PDs who had a symptom duration > 3 years at enrollment were also included for model training (see Supplementary Table 1).

PD progression subtype derivation in the PPMI cohort

Our aim was to comprehensively use consecutive clinical manifestations and CSF biomarkers during the full progression course of early PDs to identify subtypes, each of which can reflect unique progression pattern. We developed a deep learning model, named deep progression embedding (DPE) model, which learns a machine-readable representation vector for the patient based on his/her longitudinal clinical data (see Figure 1). To take full advantage of data available to enhance training of the DPE model, we used data of all included participants in the PPMI cohort, including early PDs, HCs, and SWEDDs (see Supplementary Table 1). Using the learned representation vectors of the PD patients, clustering analysis was performed and identified 3 subtypes (see Supplementary Figure 1 and Supplementary Results). Demographics and baseline clinical characteristics, 2-year follow-up clinical characteristics, and 5-year follow-up clinical characteristics can be found in Table 1 and Supplementary Tables 3 and 4, respectively. PD progression patterns in averaged trajectories and annual progression rates of diverse manifestations estimated by linear mixed effect model were illustrated in Figure 2 and Table 2, respectively.

Subtype I, mild baseline symptom and mild symptom progression, consisted of 145 (35.7%) patients. Compared to the other subtypes, Subtype I included relatively less male (N=67 [46.2%])

patients with younger age at onset (58.1±9.9 years). Subtype I showed milder motor and non-motor symptoms at baseline among the PD population (see Table 1). More importantly, in terms of symptom progression, Subtype I showed the lowest annual change rates in majority of PD symptoms within 5 years following enrollment (see Figure 2a and Table 2). More specifically, overall motor progression is low in MDS-UPDRS Part II (0.42/year, 95% CI [0.24, 0.61], *P*-value < 0.001) and III (1.04/year, 95% CI [0.64, 0.45], *P*-value < 0.001), and Schwab and England score (-0.97 /year, 95% CI [-1.25, -0.69], *P*-value < 0.001). Meanwhile, non-motor symptoms also progressed with a mild rate, e.g., MDS-UPDRS Part I only had an estimated 0.60 (95% CI [0.45, 0.75], *P*-value < 0.001) increase per year.

Subtype II, mild baseline symptom and moderate symptom progression, consisted of 207 (50.9%) patients. Subtype II included more male (N=155 [74.9%]) patients with relatively young age at onset (59.4±10.0 years). Interestingly, although Subtype II had mild motor and non-motor symptoms at baseline (see Table 2) similar to Subtype I, worse clinical symptoms at 2-year and 5-year follow-up (see Supplementary Tables 3 and 4) and worse progression rates (generally at a moderate level) in most clinical manifestations were observed in the Subtype II, compared to that in the Subtype I (see Figure 2a and Table 2). For example, in motor manifestations, MDS-UPDRS Part II showed a 1.05 (95% CI [0.91, 1.19], *P*-value < 0.001) increase per year, MDS-UPDRS Part III a 2.10 increase, and Schwab and England score a 1.74 decline (95% CI [-1.99, -1.49], *P*-value < 0.001) per year. In non-motor symptom, for instance, MDS-UPDRS Part I had an estimated annual increase of 0.91(95% CI [0.78, 1.05], *P*-value < 0.001).

Subtype III, rapid symptom progression, consisted of 54 (13.3%) patients. Subtype III included more male (N=44 [81.5%]) patients with significantly older age at onset (64.4±8.6 years). At baseline, Subtype III showed the severest motor symptoms, e.g., MDS-UPDRS Part II and III, and Schwab and England score, as well as some non-motor manifestations especially the cognition functions (see Table 2). Importantly, Subtype III was associated with the most rapid annual progression rates in majority of motor symptoms like MDS-UPDRS Part II (2.01 /year, 95% CI [1.55, 2.47], *P*-value <

0.001) and III (3.10 /year, 95% CI [2.14, 4.06], *P*-value < 0.001), and Schwab and England score (-3.95/year, 95% CI [-5.00, -2.90], *P*-value < 0.001) as well as non-motor symptoms like MDS-UPDRS Part I (1.82 /year, 95% CI [1.48, 2.16], *P*-value < 0.001) (see Figure 2a and Table 2). In addition, Sankey diagrams showed that patients of Subtype III had an increasing risk of conversion to PIGD, cognition disorders (MCI or dementia), and depression disorders (see Figure 2b-c, and Supplementary Figure 2).

Subtype validation in the PDBP cohort

By replicating the analysis on the independent validation cohort (PDBP), 3 subtypes were also identified among early PDs in the PDBP (see Supplementary Figure 3 and Supplementary Results). Demographics, baseline and follow-up (2-year and 4-year) clinical characteristics, and progression patterns in averaged trajectories and estimated annual progression rates in the clinical manifestations of the replicated subtypes showed similar patterns with the ones identified in our primary analysis in the PPMI cohort (see Supplementary Figure 4 and Supplementary Tables 5-8). These observations demonstrated reproducibility of our novel subtyping approach and the progression subtypes identified.

CSF biomarkers

We evaluated CSF biomarkers the subtypes. As shown in Table 1, we observed significant signals in baseline Aβ-42, Aβ-42/T-tau ratio, and P-tau/α-synuclein ratio levels separating the rapid progression subtype (Subtype III) from the other subtypes. In line with clinical assessments, these baseline markers can't distinguish Subtypes I and II. We didn't find significant differences in α-synuclein, P-tau, and T-tau levels across the subtypes. We further assessed annual progression rates of the CSF biomarkers based on linear mixed effect models. As shown in Table 2, Subtype III was associated with the highest annual decrease in Aβ-42 and Aβ-42/T-tau ratio among the PD population, and those

biomarkers of the Subtype II also decreased faster than that in the Subtype I. We didn't observe significant progression trends in other PD-related CSF biomarkers among the subtypes.

Imaging biomarkers

We assessed 1-year brain atrophy measured by cortical thickness and white matter volume from 34 brain region of interests (ROIs), defined by the Desikan-Killiany atlas (averaged over the left and right hemispheres). As shown in Figure 3a, ROIs of which atrophy measures significantly (P-value < 0.05) separating a pair of PD subtypes were marked in red and color density indicates the significance level (i.e., $-\log_{10}(P)$). Overall, more ROIs with white matter volume atrophy were found significantly differentiating the subtypes. Specifically, compared to Subtype I, white matter volume atrophy in ROIs including rostral middle frontal, superior frontal, middle and Inferior temporal, fusiform, parahippocampal, superior parietal, etc. and cortical thickness atrophy in posterior cingulate were significantly higher in the Subtype II. In addition, compared to Subtype II, Subtype III showed higher white matter volume atrophy in rostral middle frontal and pars triangularis as well as cortical thickness atrophy in paracentral. Last, compared to Subtype I, Subtype III gained more white matter volume atrophy in ROIs including rostral middle frontal, pars triangularis, pars opercularis, caudal middle frontal, superior frontal, paracentral, fusiform, and parahippocampal, as well as cortical thickness atrophy in paracentral.

Genetic components of subtypes

We assessed genetic components of the identified PD subtypes using hypergeometric analysis to identify genetic factors associated with each subtype[13]. First, we didn't find significant signals in APOE alleles, and GBA and LRRK2 variants across the subtypes (see Figure 3b). We examined 90 PD-related single-nucleotide polymorphisms (SNPs)[14] reported in the latest Genome wide association study SNPs were found significantly associated with each of the three subtypes (see Figure 3b).

Specifically, out of the 90 PD-related SNPs[14], we identified 12 SNPs associated with the rapid progression subtype (Subtype III), including rs10797576, rs10847864, rs112485576, rs11683001, rs12528068, rs12951632, rs1941685, rs2248244, rs35749011, rs55961674, rs76763715, and rs997368; 15 SNPs associated with the mild progression subtype (Subtype I), including rs7938782, rs77351827, rs896435, rs2086641, rs6497339, rs76904798, rs6658353, rs13294100, rs10847864, rs9261484, rs10748818, rs11150601, rs76949143, rs2042477, and rs850738; and 15 SNPs associated with the mild baseline but moderate progression subtype (Subtype II). Last, the subtypes showed similar distribution of GRS derived by the 90 loci (see Table 2).

Predictive model of PD progression subtypes

The identified subtypes captured long term symptom progression trends of PD. To enhance their usability in clinical practice like prognosis prediction and participants stratification in clinical trial enrollment, there is the need of prediction of the identified PD progression subtypes short-term information after baseline (e.g., within 1-year following baseline). To this end, we built predictive model of PD progression subtypes using patient data including demographics, baseline and 1-year follow-up clinical assessments and CSF biomarkers, 10 principal components derived from 1-year brain atrophy, and genetic data. We built the predictive model using a cascade classifier consisting of two Random Forests: one distinguishes Subtype III from the others and the other distinguishes Subtype I and Subtype II. As shown in Figure 4a, the predictive model can separate Subtype III from the others very well, with an area under receiver operating characteristics curve (AUC) of 0.87 ± 0.05. Distinguishing Subtypes I and II from each other is difficult due to that the two subtypes have similar baseline clinical characteristics, yet the model still achieved a good performance with an AUC of 0.74 ± 0.07. Last, the novel SHapley Additive exPlanation (SHAP) values[15] were calculated to demonstrate contributions of those predictors in separating the subtypes (see Figure 4b-c). Motor assessments like MDS-UPDRS Part II and III and Schwab and England score, non-motor assessments like semantic fluency and Symbol–Digit Matching (SDM), CSF biomarkers like P-tau/α-synuclein

ratio, and genetic variants like rs10797576 played a relatively important role in separating Subtype III from the others. Non-motor assessments like semantic fluency, letter-number sequencing, and MoCA, CSF biomarkers like P-tau/α-synuclein ratio, genetic variants like rs76904798, imaging components, and demographics were more important in distinguishing Subtypes I and II.

Discussions

Identification of PD subtypes is crucial in understanding clinical and pathological mechanisms of the complex disease, which may accelerate disease-modification treatment development towards precision medicine. Challenges that hindered recent PD subtyping studies included symptom heterogeneity[4,5] and progression heterogeneity[5]. To address these, data-driven approaches have led to a novel route of comprehensively integrating patient clinical data to identify PD subtypes[7,16]. Early efforts typically performed clustering analysis either on clinical data collected at baseline[17,18] or on concatenation of data collected at baseline and follow-up time points to capture symptom progression[19-21]. Though these identified subtypes showed prognostic implications, the full course progression trajectories of patients were insufficiently measured and hence the underlying progression patterns were probably not well modeled. Novel data-driven technique that can model the high dimensional and longitudinal data to capture underlying evolving patterns are urgently needed. We[12] previously utilized the LSTM model, a deep learning model specifically designed for multivariate time sequence data modeling task. The model was trained in a supervised manner, where we need to specify a set of target features as supervision information for model training, which is actually subjective and hence may limit generalizability and usability of the model. In the present study, we improved our LSTM model and then proposed the DPE (deep progression embedding) model to model patient symptom progression data for PD subtyping. Compared to previous PD subtyping efforts, contributions of this investigation included: 1) the DPE used LSTM units to construct a deep autoencoder, which enabled us to comprehensively model patients' symptom progression in diverse clinical manifestations and CSF biomarkers within the full course of PD, but also to train model in an

unsupervised manner without any assumptions. 2) Using patients' representation vectors produced by DPE, encoded with PD progression patterns, we successfully identified three PD subtypes in the PPMI cohort. Further, replication of our method in the independent cohort, i.e., PDBP, confirmed reproductivity of our approach and the three subtypes. 3) CSF biomarkers, neuroimaging biomarkers, genetic components of the subtypes suggested potential biological pathways that lead to distinct PD progression patterns and predictive model was trained to distinguish the progression subtypes within 1-year following baseline.

Importantly, each of the three subtypes showed significant progression pattern of PD. Subtype III (the rapid symptom progression subtype) had the worst symptoms in most motor and non-motor manifestations at baseline and 2-year and 5-year follow-up. This is similar to the "diffuse malignant" subtype reported in the previous study[20] but describes a sub-population with significant symptom progression rates that adds more prognostic implications. Specifically, compared to the "diffuse malignant" subtype, our identified Subtype III had relatively better overall symptom at baseline. Second but more important, in most motor and non-motor manifestations of PD, Subtype III showed the worst annual progression rates estimated by the linear mixed effects models among all PDs (see Figure 2a). In line with the clinical manifestation trajectories, Subtype III also showed steady pathways of clinical phenotype evolution. For instance, in motor phenotypes, an increasing proportion of patients developed to PIGD while the proportion of tremor dominant shrank (see Figure 2b). Similarly, patients of Subtype III also had a clear increasing risk to suffer from cognition dysfunction like dementia and depression symptoms over time (see Figure 2c and Supplementary Figure 2). Subtype III potentially represents the sub-population that suffered from the fastest functional deterioration in PD. In addition, Subtype I (the mild baseline symptom and mild symptom progression subtype) and Subtype II (the mild baseline symptom and moderate symptom progression) had similar baseline symptom severities (see Table 1), however, at 2-year follow-up significant differences emerged in their motor manifestations and at 5-year follow-up they significantly differed in both motor and non-motor manifestations (see Supplementary Tables 3 and 4). Moreover, Subtype II had higher annual progression rates in most clinical manifestations compared to that of the Subtype I.

More than intuitively "worse baseline leading to worse follow-up while milder baseline resulting in better follow-up", our data highlight that patients with similar baseline severity can have distinct progression patterns and develop distinct prognosis. The findings emphasized the importance of identifying progression subtypes of PD.

As the subtypes described distinct symptom progression patterns during the course of PD, identifying demographic factors, early biomarkers, and genetic factors may add significant prognostic implications. First, demographics including age at onset and gender were potential factors of the subtypes. Specifically, Subtypes I and II had a younger age at onset compared to that of the Subtype III, in line with the previous findings that late onset PDs associated with severer clinical symptoms[22,23]. Male was associated with Subtypes II and III, which showed worse PD progression rates, in accordance with previous progression factor findings[24,25]. From the univariate perspective, motor (such as MDS-UPDRS Part II and III, Schwab and England score and PIGD score) and cognition symptoms (such as MoCA, Benton judgment of line orientation, HVLT, LNS, semantic fluency, and symbol digit test) were potential factors to separate Subtype III from Subtypes I and II at baseline. Limited signal in baseline clinical symptoms were found as factors of Subtypes I or II.

In CSF biomarkers, decreased Aβ-42 and Aβ-42/T-tau ratio, which have been associated with cognition impairment and progression of cognition symptoms in PD[26,27], were found as potential biomarkers of Subtype III at baseline. We observed significant decreasing trends in the two biomarkers in the Subtype III. This would be a reason that Subtype III suffered from continuously increasing burden of MCI and dementia. Moreover, this may also suggest cognition impairment induced by disturbed Aβ metabolism as a potential pathological pathway in the rapid progression subtype of PD (Subtype III). Baseline Aβ-42 level and Aβ-42/T-tau ratio were not significantly different between Subtypes I and II, but in line with their clinical symptom progression trends, Subtype II was associated with relatively worse annual change rates in the two biomarkers. In addition, baseline P-tau/α-synuclein ratio was also found as a potential biomarker for Subtype III but no further evidence was found in its progression trajectory over time.

Using serial MRI data, brain atrophy was measured to identify biomarkers in PD. Atrophy in brain cortical thickness has previously been associated with disease stage[28,29], cognition impairment[30], and even dementia[28] in PD. In addition, signals of reduction in brain white matter have also been reported to correlate to motor and cognition dysfunctions[31,32], as well as their progression[33]. In the present investigation, we discovered interesting patterns of brain atrophy in brain cortical thickness and white matter volume within 1-year follow-up in separating the PD progression subtypes, suggesting potential biomarkers and brain atrophy pathways. Overall, more signals were found in the regional white matter volume reduction compared to those in the regional cortical thinning. White matter volume reduction in regions of frontal lobes like rostral middle frontal, caudal middle frontal, superior frontal, etc., which are usually involved in motor function, memory, language, etc., were found to consistently distinguish each of the three subtypes from others. The frontal lobe related pathways may play a crucial rule in PD progression. White matter volume reduction in temporal lobes, which mainly correlate to memory, language understanding, and other non-motor functions, were found as potential indicators in differentiating the mild progression subtype (Subtype I) with others. In addition, cortical thinning related pathology in the paracentral lobe probably plays a role in the mechanism of rapid symptom progression.

We also studied genetic components of the progression subtypes from different levels: the variants in PD or neurodegeneration related genes including APOE, GBA, and LRRK2, the SNPs of PD-related risk loci[14], and GRS derived from these loci. Due to the low prevalence of GBA and LRRK2 risk variants in the PD population in the PPMI cohort, P-values didn't show significant signals in these two genes. APOE $\varepsilon 4$, which plays a role in neurodegenerative diseases like Alzheimer's disease and dementia with Lewy bodies, was likely to not contribute to the progression subtypes of PD. Interestingly, though GRS distributed similarly among the three subtypes, different PD risk loci were associated with the subtypes, indicating distinct genetic components resulting in different progression patterns in PD. In this point, further investigation is needed to explore underlying pathological pathways of these loci in the future.

Apart from the exploration of the factors and biomarkers in the univariate manner, we performed predictive modeling of the subtypes using data within 1-year follow-up. Our model comprehensively considered a wide range of candidate predictors including demographic, clinical, and genetic factors, as well as CSF and imaging biomarkers. The predictive model, being able to distinguish the three progression subtypes well, adds clinical implications to this study, including: 1) even though each of the subtypes presenting unique progression pattern and follow-up symptoms, one can predict prognosis early (within 1-year follow-up); 2) it provides a feasible and accurate way to apply the identified subtypes to clinical practice or trial enrollment, as it can be used to predict subtype membership of new patients with limited longitudinal data. In addition, contributions of the predictors calculated by the SHAP method showed interesting patterns. Specifically, motor and non-motor characteristics play a more important role in predicting Subtype III out of Subtypes I and II. In contrast, cognition manifestations combined with genetic risk factor, CSF biomarkers, and neuroimaging components have a higher contribution on separating Subtypes I and II early. In total, these signals were in line with the univariate findings.

While this study presents a new contribution in the efforts to parse the symptom heterogeneity and progression heterogeneity in PD, there remain several limitations. First of all, our novel approach that captures PD progression patterns for subtype identification is completely data-driven and relies on data availability and data quality. PPMI and PDBP collect longitudinal records in a board spectrum of clinical assessments and biomarkers of patients, which enable us to model the complex symptom progression trajectories of PD to identify subtypes. Sample sizes of the cohorts are relatively large, while from the data-driven perspective, it's always more data the better especially in model training, as abundant data can help avoid model underfitting and overfitting and hence lead to the model capturing true underlying patterns from data[8,9,34]. Meanwhile, data quality, especially missing values may affect the robustness of the identified subtypes. To address this issue, we excluded patients who had insufficient longitudinal information (less than 1-year records) and used a combination of the last observation carried forward (LOCF) and next observation carried backward (NOCB) methods[35] to

impute missing values while retaining longitudinal consistency. The variables missing all values along timeline of a patient were imputed by median values of the population. Even so, we still missed these real values hence may incorporate bias. Second, although the use of deep learning model, the DPE, demonstrated its feasibility in modeling the high-dimensional, longitudinal clinical records of patient to uncover disease progression patterns and successfully identified PD subtypes with significant progression patterns, the model interpretability, a common concern of deep learning approaches[8,9,34], could be an issue that limits utilization of our method in clinical practice. Third, we assessed neuroimaging biomarkers within 1-year follow-up and genetic factors aiming at separating the long-term progression subtypes early. This also makes the present study unique compared to previous PD progression subtyping efforts[12,21]. Though preliminary evidence has been found in this study, there remain space. For instance, fine-grained (voxel-level) signals from neuroimages and whole genome sequencing data may add more insights. How these biomarkers and factors interact to construct biological pathways that finally result in distinct progression patterns of PD is also unclear. In this case, causal inference on the top of our subtypes may help.

Methods

Study design and cohort description

The present study included two longitudinal PD cohorts for analysis: the Parkinson's Progression Markers Initiative (PPMI, http://www.ppmi-info.org)[10] and the Parkinson Disease Biomarkers Program (PDBP)[11].

PPMI, launched in 2010 and sponsored by the Michael J. Fox Foundation, is an international and multi-center observational study that longitudinally and comprehensively collects data and bio-samples from *de novo* PDs (PD diagnosis within the last 2 years and untreated at enrollment), healthy

control (HC) volunteers and at-risk populations. More information about PPMI participants has been described elsewhere[10]. The institutional review board approved the PPMI protocol in all participating sites. Data were obtained from PPMI website (https://www.ppmi-info.org) under PPMI Data Use Agreement.

PDBP, established in 2012 and funded by the National Institute of Neurological Disorders and Stroke (NINDS), is an observational study that longitudinally and comprehensively collects data and bio-samples from patients with PDs and HC participants. Compared to PPMI, PD participants in PDBP have a broad distribution of stages and duration of PD. More information about PPMI participants has been described elsewhere[11]. The institutional review board approved the PDBP protocol in all participating sites. Data were obtained via the Accelerating Medicines Partnership Parkinson's disease (AMP-PD) platform (http://amp-pd.org) under AMP-PD Data Use Agreement.

Overall workflow of the present study is illustrated in Figure 1a. We used PPMI as the development cohort. We included HCs, subjects with scans without evidence for dopaminergic deficit (SWEDD), and *de novo* treatment-naïve PD patients to model symptom progression pattern to produce patient representation vectors. Patients who had less than 1-year historical records were excluded as lack of longitudinal information. Subtypes were identified using the learned patient representation vectors among the PDs. In addition, PDBP was used as the independent cohort for replication of our identified PD subtypes. Similarly, we included all HCs and PDs in the PDBP to model symptom progression pattern to produce patient representation vectors and excluded patients who had less than 1-year historical records. As our aim was to model progression patterns of early PD, in the PDBP, only the early PDs (symptom duration < 3 years at enrollment) were used to identify PD subtypes. More details of the studied cohorts and data utilization were illustrated in the Supplementary Table 1.

Clinical variables for subtyping

PPMI comprehensively and longitudinally collects patient's clinical assessments including both motor and non-motor manifestations and CSF biospecimen examination data. In this study, we used and extracted motor manifestation data including Movement Disorders Society–revised Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part II and III[36] and Schwab-England activities of daily living score, as well as non-motor manifestation data including MDS-UPDRS Part I[36], Scales for Outcomes in Parkinson's disease-Autonomic (SCOPA-AUT)[37], Geriatric depression scale (GDS)[38], Questionnaire for Impulsive-Compulsive Disorders in Parkinson's disease (QUIP)[39], State-Trait Anxiety Inventory (STAI)[40], Benton Judgment of Line Orientation (JOLO)[41], Hopkins Verbal Learning Test (HVLT)[42], Letter-number sequencing (LNS), Montreal Cognitive Assessment (MoCA)[43], Semantic verbal-language fluency test[44], Symbol–Digit Matching (SDM)[45], Epworth Sleepiness Score (ESS)[46], REM sleep behaviour disorder (RBD)[47], and Cranial Nerve Examination. CSF biospecimen examination data were also analyzed. We collected α-synuclein measured by enzyme-linked immunosorbent assay[48], and amyloid-beta1–42 ($\alpha\beta$-42), phosphorylated Tau protein at threonine 181 (P-tau$_{181}$), and total tau protein (T-tau) measured by INNO-BIA AlzBio3 immunoassay.

For the PDBP cohort, we used the clinical variables shared with the PPMI cohort. More details of the clinical variables used for PD subtyping can be found in the Supplementary Table 2.

Data preparation

To identify PD progression subtypes that capture symptom progression pattern during the course of the complex disease, we used longitudinal clinical assessment and CSF biomarker data of patients for PD subtyping. In the PPMI cohort, we extracted data at baseline and up-to 12 regular follow-up visits (5 years) for each patient. Baseline visit and Visits 1-4 were regularly scheduled every 3 months while other visits every 6 months. In this way, each patient was associated with a multivariate clinical time

sequence (see Figure 1a). After that, missing values were imputed using the last observation carried forward (LOCF) and next observation carried backward (NOCB) strategies[35]. The variables missing all values along timeline of a patient were imputed by median values of the population. In order to eliminate the effects of value magnitude, all variables were scaled based on z-score.

Identification of PD progression subtypes

Learning progression embedding vectors of patients using deep learning

Our aim was to identify PD subtypes, each of which can reflect a unique PD symptom progression pattern within the course of the disease. To this end, there is the need of fully considering longitudinal data of patients to derive PD subtypes. In this investigation, we developed a deep learning model, named deep progression embedding (DPE), which takes the multivariate clinical time sequence data of each patient as input to learn a machine-readable vector representation for the patient while encoding his/her PD progression pattern over time (see Figure 1b). Our DPE model was based on the Long-Short Term Memory (LSTM) units[8,49], a kind of deep learning model typically developed to learn vector representations for time sequence data. Using a "memory cell", the LSTM can maintain historical information in memory for long periods of time, and hence is a good fit of modeling disease progression based on patient's longitudinal clinical data[50]. Details of mechanism of the LSTM can be found in the Supplementary Methods. In addition, we used an autoencoder architecture that allows us to train the DPE model in an unsupervised manner. It's consisted of two components, an encoder and a decoder, each of which is a LSTM model: 1) the encoder takes multivariate clinical time sequence of each patient as input and learn a vector representation that preserves his/her PD progression; 2) while the decoder tries to reconstruct input time sequence of the patient. Then, the model was trained by minimizing difference between the input multivariate clinical time sequences and the reconstructed ones. To take full advantage of data available to enhance model training, we used data of all included patients, including PDs, HCs, and SWEDDs, in the PPMI cohorts. After the DPE was trained, we obtained a learned vector representation for each patient, representing his/her PD progression pattern.

Identification of PD progression subtypes

Next, we performed clustering analysis to identify PD progression subtypes in the PPMI cohort. More specifically, agglomerative hierarchical clustering with Euclidean distance calculation and Ward linkage criterion[51] was applied to the DPE learned vector representations that have undergone dimension reduction by Student t-Distributed Stochastic Neighbor Embedding (t-SNE)[52]. To determine the optimal number of clusters (i.e., subtypes), we used the dendrogram naturally produced by the agglomerative hierarchical clustering algorithm as well as 18 measures of clustering models provided by 'NbClust' software[53]. More details can be found in the Supplementary Methods.

Subtype validation in the PDBP cohort

To enhance model reproductivity, we validated the subtypes in the PDBP cohort. We repeated the whole procedure above to replicate the subtypes. We re-trained DPE model to calculate patients' vector representations, re-do clustering analysis for subtyping, and characterized subtypes. To re-train DPE, we used data of all included patients in PDBP, including all PDs and HCs. PD participants in PDBP have a broad distribution of PD duration of PD[11]. In accordance with our primary analysis in the PPMI cohort, we focused on the early PD population (whose PD duration < 3 years) and performed agglomerative hierarchical clustering to re-derive the PD progression subtypes.

Statistical methods

Subtype characterization

We characterize the identified subtypes in two ways. First, we characterized the subtypes by evaluating their differences in demographics and all clinical assessments and CSF biomarker levels at baseline and 2- and 5-year follow-up. For group comparisons, we performed analysis of variance

(ANOVA) for continuous data and χ^2 test for categorical data. Analysis of covariance (ANCOVA) was also applied, adjusting for age at symptom and sex. Two-tailed P-values smaller than 0.05 were considered as the threshold for statistical significance.

Second, we evaluated PD progression rates in terms of annual changes in motor and non-motor clinical manifestations and CSF biomarkers by subtypes. More specifically, for each subtype, we fitted a linear mixed effects model for each clinical manifestation or CSF biomarker to estimate its progression rate (annual change), adjusting for age at symptom and sex. We also evaluated the conversion trends of motor phenotypes (tremor dominant and postural instability and gait disorder [PIGD]) as well as non-motor phenotypes including cognition (normal cognition, mild cognition impairment [MCI], dementia), REM sleep behavior disorder (RBD [negative or positive]), and depression (normal, mild, moderate, and severe).

Neuroimaging biomarkers

High resolution T1-weighted 3 T Magnetic Resonance Imaging (MRI) data of patients were available at baseline and follow-up in PPMI. We analyzed patients' longitudinal MRIs at baseline and 1-year follow-up. Specifically, for each patient, we calculated 1-year brain atrophy measured by cortical thickness and white matter volume from 34 region of interests (ROIs), defined by the Desikan-Killiany atlas (averaged over the left and right hemispheres). We evaluated those measures in separating each pair of PD subtypes. The Student's t-tests were used for two-group comparisons. The 'ggseg'[54] in R was used to visualize the neuroimaging biomarkers under the Desikan-Killiany atlas.

Genetic analysis

To evaluate difference of genetic components among PD subtypes, we analyzed genetic data in the PPMI cohort. APOE $\varepsilon 2$ and $\varepsilon 4$ alleles, GBA and LRRK2 variants, and genotypes of 90 PD-related risk loci[14] were collected for analysis. We also used genetic risk score (GRS) derived from the 90

single-nucleotide polymorphisms (SNPs), showing cumulative burden of the risk of individuals. For each genetic factor, e.g., GBA variant or SNP, the hypergeometric analysis was performed to identify SNPs that were enriched in each subtype out of the PD population. *P* value from the hypergeometric analysis indicates association of a SNP to the specific subtype[13].

Early prediction of PD progression subtypes

To enhance their usability in clinical practice, there is the need of early prediction of the identified PD progression subtypes. To this end, we built predictive model of subtypes using information collected within the first year since enrollment. Specifically, we used demographics, 10 principal components derived from neuroimaging biomarkers (1-year brain atrophy measured by cortical thickness and white matter volume from 34 ROIs), 90 PD-related SNPs, and clinical variables at baseline and 1-year follow-up as candidate predictors. To improve model robustness and practicability in separating multiple subtypes, we built the predictive using cascade classifier[55], which was proposed to address multi-label classification tasks. Specifically, it contained a sequential of multiple basic classifiers, such that in each step, it predicted a subtype from the remaining subjects that will be sent to the successor basic classifier. Here we used Random Forest as the basic classifier and the model was trained using a 5-fold cross-validation strategy. We used the receiver operating characteristics curve (ROC) and area under ROC curve (AUC) to measure prediction performance. The novel SHapley Additive exPlanation (SHAP) values[15] were used to assess contributions of the predictors in distinguishing each subtype from the others. More specifically, SHAP value shows distribution of the impact each feature has on the model output.

Data availability

Data from PPMI can be download from the official web site (http://www.ppmi-info.org) via request. Data from the PDBP can be download from AMP-PD platform (http://amp-pd.org) via request.

Code availability

All computer codes in this study are available at https://github.com/changsu10/Parkinson_progression_subtyping. Implementation of our work is based on Python 3.6 and R 3.6. More specifically, deep learning model was implemented based on Python package PyTorch 1.1.0 (https://pytorch.org). Data extraction and preparation were performed using Python packages Pandas 1.2.1 (https://pandas.pydata.org) and NumPy 1.20.1 (https://numpy.org). tSNE, agglomerative hierarchical clustering model, and Random Forest classifier were implemented using Python packages scikit-learn 0.23.2 (https://scikit-learn.org/stable/) and scipy 1.5.3 (https://www.scipy.org). SHAP values were calculated using Python package SHAP 0.35.0 (https://shap.readthedocs.io/en/latest/). R package 'NbClust' (https://cran.r-project.org/web/packages/NbClust/NbClust.pdf) was used to calculate measures of clusters to determine the optimal cluster number in agglomerative hierarchical clustering. Linear mixed effects models were performed using R package lme4 (https://cran.r-project.org/web/packages/lme4/index.html). All other statistical tests were performed based on R.

Author contributions

FW and CS for conceptualization, investigation, model design, writing, reviewing and editing of the manuscript. CS, YH, and MB for data analysis. CH for discussion, commenting and editing the manuscript.

Competing interests

The authors have declared that no conflict of interest exists.

References

1 Poewe, W. *et al.* Parkinson disease. *Nature Reviews Disease Primers* **3**, 17013 (2017).

2 Balestrino, R. & Schapira, A. H. V. Parkinson disease. *European Journal of Neurology* **27**, 27-42 (2020).

3 Mizuno, Y. Where do we stand in the treatment of Parkinson's disease? *Journal of Neurology* **254**, 13-18 (2007).

4 Lewis, S. J. G. *et al.* Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. *Journal of Neurology, Neurosurgery & Psychiatry* **76**, 343 (2005).

5 Greenland, J. C., Williams-Gray, C. H. & Barker, R. A. The clinical heterogeneity of Parkinson's disease and its therapeutic implications. *European Journal of Neuroscience* **49**, 328-338 (2019).

6 van Rooden, S. M. *et al.* The identification of Parkinson's disease subtypes using cluster analysis: A systematic review. *Movement Disorders* **25**, 969-978 (2010).

7 Fereshtehnejad, S.-M. & Postuma, R. B. Subtypes of Parkinson's disease: what do they tell us about disease progression? *Current neurology and neuroscience reports* **17**, 34 (2017).

8 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).

9 Si, Y. *et al.* Deep representation learning of patient data from Electronic Health Records (EHR): a systematic review. *arXiv preprint arXiv:2010.02809* (2020).

10 Marek, K. *et al.* The Parkinson Progression Marker Initiative (PPMI). *Progress in Neurobiology* **95**, 629-635 (2011).

11 Rosenthal, L. S. *et al.* The NINDS Parkinson's disease biomarkers program. *Movement Disorders* **31**, 915-923 (2016).

12 Zhang, X. *et al.* Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study. *Scientific Reports* **9**, 797 (2019).

13 Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine* **7**, 311ra174 (2015).

14 Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology* **18**, 1091-1102 (2019).

15 Lundberg, S. M. & Lee, S.-I. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (Curran Associates Inc., Long Beach, California, USA, 2017).

16 Qian, E. & Huang, Y. Subtyping of Parkinson's Disease - Where Are We Up To? *Aging Dis* **10**, 1130-1139 (2019).

17 Erro, R. *et al.* The Heterogeneity of Early Parkinson's Disease: A Cluster Analysis on Newly Diagnosed Untreated Patients. *PLOS ONE* **8**, e70244 (2013).

18 Erro, R. *et al.* Clinical clusters and dopaminergic dysfunction in de-novo Parkinson disease. *Parkinsonism & Related Disorders* **28**, 137-140 (2016).

19 Fereshtehnejad, S.-M. *et al.* New Clinical Subtypes of Parkinson Disease and Their Longitudinal Progression: A Prospective Cohort Comparison With Other Phenotypes. *JAMA Neurology* **72**, 863-873 (2015).

20 Fereshtehnejad, S.-M., Zeighami, Y., Dagher, A. & Postuma, R. B. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. *Brain* **140**, 1959-1976 (2017).

21 Faghri, F. *et al.* Predicting onset, progression, and clinical subtypes of Parkinson disease using machine learning. *bioRxiv*, 338913 (2018).

22 Diederich, N. J., Moore, C. G., Leurgans, S. E., Chmura, T. A. & Goetz, C. G. Parkinson Disease With Old-Age Onset: A Comparative Study With Subjects With Middle-Age Onset. *Archives of Neurology* **60**, 529-533 (2003).

23 Pagano, G., Ferrara, N., Brooks, D. J. & Pavese, N. Age at onset and Parkinson disease phenotype. *Neurology* **86**, 1400 (2016).

24 Reinoso, G. *et al.* Clinical evolution of Parkinson's disease and prognostic factors affecting motor progression: 9-year follow-up study. *European Journal of Neurology* **22**, 457-463 (2015).

25 Pigott, K. *et al.* Longitudinal study of normal cognition in Parkinson disease. *Neurology* **85**, 1276 (2015).

26 Kang, J.-H. *et al.* CSF biomarkers associated with disease heterogeneity in early Parkinson's disease: the Parkinson's Progression Markers Initiative study. *Acta Neuropathologica* **131**, 935-949 (2016).

27 Terrelonge, M., Marder, K. S., Weintraub, D. & Alcalay, R. N. CSF β-Amyloid 1-42 Predicts Progression to Cognitive Impairment in Newly Diagnosed Parkinson Disease. *Journal of Molecular Neuroscience* **58**, 88-92 (2016).
28 Zarei, M. *et al.* Cortical thinning is associated with disease stages and dementia in Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry* **84**, 875 (2013).
29 Gao, Y. *et al.* Changes in Cortical Thickness in Patients With Early Parkinson's Disease at Different Hoehn and Yahr Stages. *Frontiers in Human Neuroscience* **12** (2018).
30 Segura, B. *et al.* Cortical thinning associated with mild cognitive impairment in Parkinson's disease. *Movement Disorders* **29**, 1495-1503 (2014).
31 Wen, M.-C. *et al.* Longitudinal brain volumetric changes and their predictive effects on cognition among cognitively asymptomatic patients with Parkinson's disease. *Parkinsonism & Related Disorders* **21**, 483-488 (2015).
32 Uribe, C. *et al.* Cortical atrophy patterns in early Parkinson's disease patients using hierarchical cluster analysis. *Parkinsonism & Related Disorders* **50**, 3-9 (2018).
33 Lee, S. H. *et al.* Regional Volume Analysis of the Parkinson Disease Brain in Early Disease Stage: Gray Matter, White Matter, Striatum, and Thalamus. *American Journal of Neuroradiology* **32**, 682 (2011).
34 Su, C., Xu, Z., Pathak, J. & Wang, F. Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry* **10**, 116 (2020).
35 Engels, J. M. & Diehr, P. Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology* **56**, 968-976 (2003).
36 Goetz, C. G. *et al.* Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders* **23**, 2129-2170 (2008).
37 Visser, M., Marinus, J., Stiggelbout, A. M. & Van Hilten, J. J. Assessment of autonomic dysfunction in Parkinson's disease: The SCOPA-AUT. *Movement Disorders* **19**, 1306-1312 (2004).
38 Sheikh, J. I. & Yesavage, J. A. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. *Clinical Gerontologist: The Journal of Aging and Mental Health* (1986).
39 Weintraub, D. *et al.* Validation of the questionnaire for impulsive-compulsive disorders in Parkinson's disease. *Movement Disorders* **24**, 1461-1467 (2009).
40 Spielberger, C. D. State-Trait Anxiety Inventory. In *The Corsini Encyclopedia of Psychology* 1-1.
41 Benton, A. L., Varney, N. R. & Hamsher, K. d. Visuospatial Judgment: A Clinical Test. *Archives of Neurology* **35**, 364-367 (1978).
42 Shapiro, A. M., Benedict, R. H. B., Schretlen, D. & Brandt, J. Construct and Concurrent Validity of the Hopkins Verbal Learning Test – Revised. *The Clinical Neuropsychologist* **13**, 348-358 (1999).
43 Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society* **53**, 695-699 (2005).
44 Gladsjo, J. A. *et al.* Norms for Letter and Category Fluency: Demographic Corrections for Age, Education, and Ethnicity. *Assessment* **6**, 147-178 (1999).
45 Smith, A. *Symbol digit modalities test*. (Western Psychological Services Los Angeles, 1973).
46 Johns, M. W. A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale. *Sleep* **14**, 540-545 (1991).
47 Stiasny-Kolster, K. *et al.* The REM sleep behavior disorder screening questionnaire—A new diagnostic instrument. *Movement Disorders* **22**, 2386-2393 (2007).
48 Kang, J.-H. *et al.* Association of Cerebrospinal Fluid β-Amyloid 1-42, T-tau, P-tau181, and α-Synuclein Levels With Clinical Features of Drug-Naive Patients With Early Parkinson Disease. *JAMA Neurology* **70**, 1277-1287 (2013).
49 Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9**, 1735-1780 (1997).
50 Ayala Solares, J. R. *et al.* Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics* **101**, 103337 (2020).
51 Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification* **31**, 274-295 (2014).
52 Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).
53 Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *2014* **61**, 36 (2014).
54 Mowinckel, A. M. & Vidal-Piñeiro, D. Visualization of Brain Statistics With R Packages ggseg and ggseg3d. *Advances in Methods and Practices in Psychological Science* **3**, 466-483 (2020).
55 Gama, J. & Brazdil, P. Cascade Generalization. *Machine Learning* **41**, 315-343 (2000).

Figures

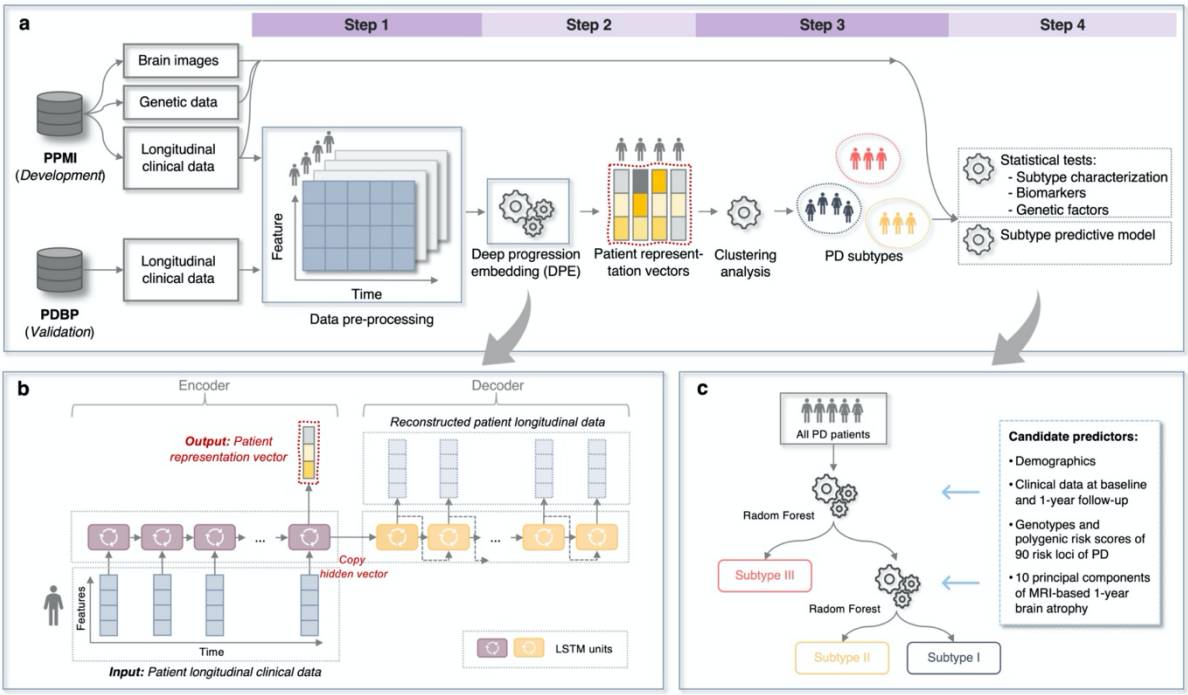

Figure 1. A schematic of the present analysis. a. Study pipeline. The PPMI and PDBP cohorts were used as the development and independent validation cohorts, respectively. First, in data pre-processing, each patient longitudinal clinical data can be represented by a multivariate time sequence of which the two dimensions represent time and features (including clinical assessments and CSF biomarkers); missing values were also imputed. Second, the deep progression embedding (DPE) model was used to learn a representation vector for each patient according to his/her longitudinal clinical data (**b**). Specifically, the DPE uses an autoencoder architecture consisting of an encoder and a decoder: 1) the encoder uses the LSTM to model patient longitudinal clinical data and produces a hidden vector; 2) the decoder uses LSTM to decompresses the hidden vector to reconstruct the input data; 3) DPE can be trained by minimize difference between the input data and reconstructed data and the final output hidden vector is the representation vector for the patient, encoding his/her symptom progression profile. Third, agglomerative hierarchical clustering was performed based on patients' representation vectors to identify subtypes. Fourth, (**c**) statistical analysis was performed to identify demographic factors, biomarkers, and genetic factors of the subtypes, and predictive model of subtype

was built based on a cascade classifier that includes two Random Forests: one distinguishes Subtype III from the others and the other distinguishes Subtype I and Subtype II.

Abbreviations: LSTM = the Long-Short Term Memory model; PD = Parkinson's disease; PDBP = the Parkinson Disease Biomarkers Program; PPMI = the Parkinson's Progression Markers Initiative.

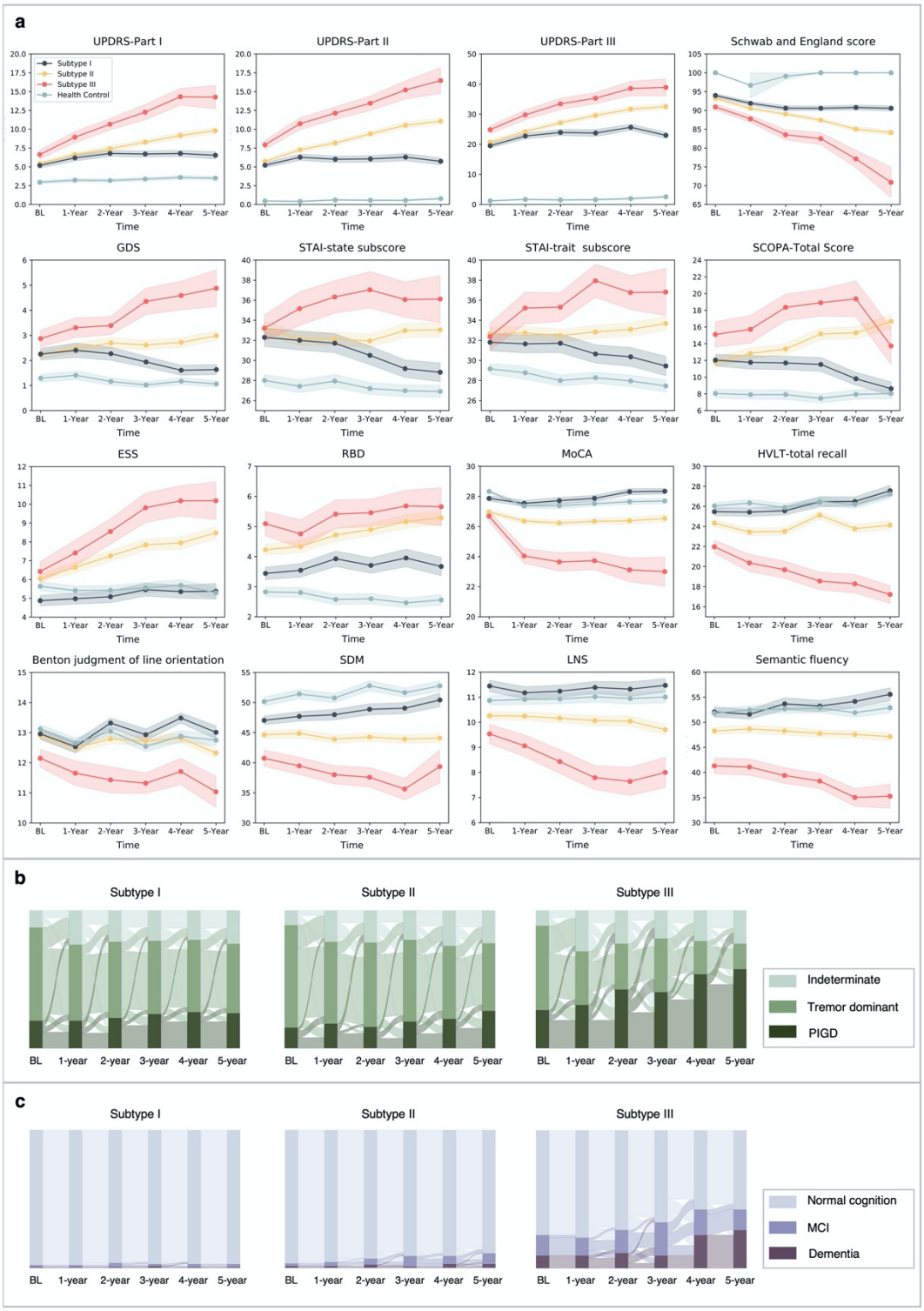

Figure 2. Progression patterns of the three PD subtypes. a. Averaged progression trajectories in clinical manifestations by subtypes. **b. and c.** Sankey diagram showing evolution patterns of motor and cognition phenotypes by subtypes.

Abbreviations: ESS = Epworth sleepiness score; GDS = Geriatric depression scale; HVLT = Hopkins Verbal Learning Test; LNS = Letter-number sequencing; MCI = mild cognition impairment; MDS-UPDRS = Movement Disorders Society–revised Unified Parkinson's Disease Rating Scale; MoCA = Montreal Cognitive Assessment; PIGD = postural instability and gait disorder; PPMI = the Parkinson's Progression Markers Initiative; RBD = REM sleep behavior disorder; SCOPA = Scales for Outcomes in Parkinson's Disease; SDM = Symbol–Digit Matching; STAI = State-Trait Anxiety Inventory.

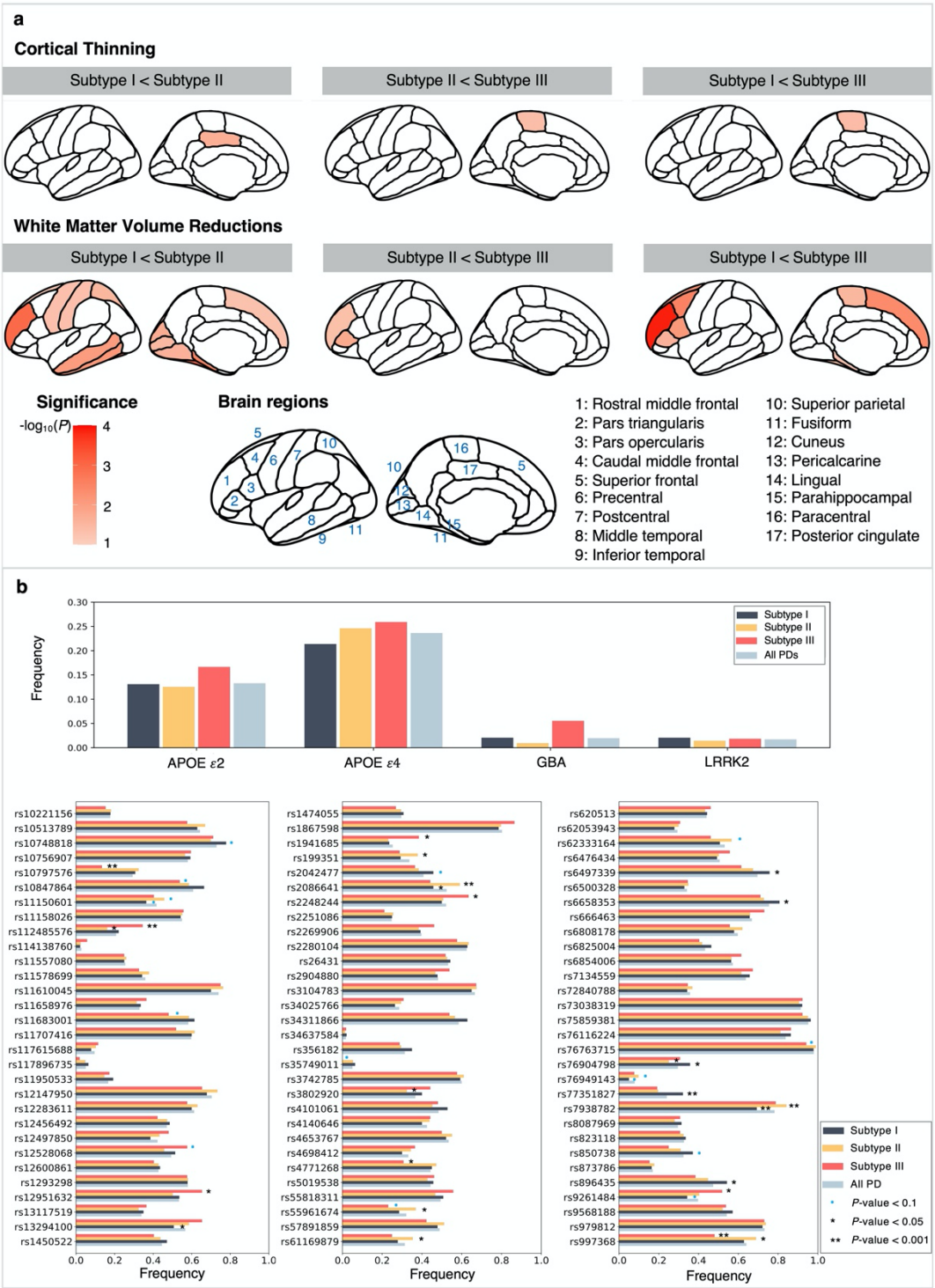

Figure 3. Neuroimaging biomarkers and genetic factors of subtypes. a. Regions showing significant signals in 1-year brain atrophy between a pair of subtypes. 1-year brain atrophy was measured by cortical thickness and white matter volume from 34 region of interests (ROIs), defined

by the Desikan-Killiany atlas (averaged over the left and right hemispheres). Color density denotes significance in terms of $-\log_{10}(P)$. **b.** Comparisons of genetic factors among subtypes.

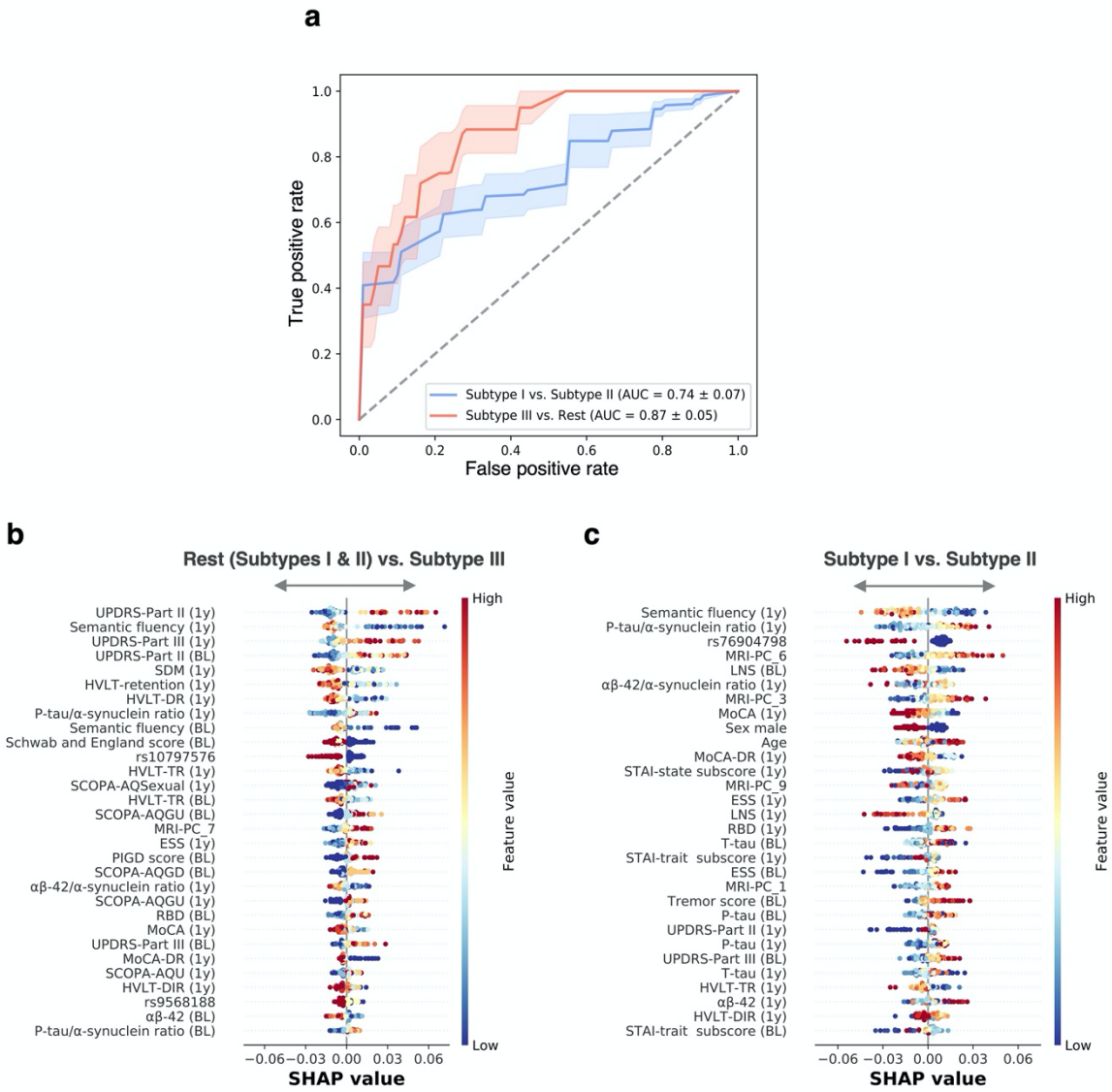

Figure 4. Predictive model of progression subtype. The predictive model was trained to predict subtype memberships of patients using baseline and 1-year follow-up data. **a.** Prediction performance in terms of used the receiver operating characteristics curve (ROC) and area under ROC curve (AUC) to measure prediction performance. An AUC score close to 1 means ideal prediction performance. **b. and c.** SHAP value-based predictor contribution visualization of the predictive model. X axis denotes SHAP value and color density indicates feature value. SHAP value shows distribution of the impact each feature has on the model output. For example, in (**b**), higher value of MDS-UPDRS Part II (at 1-year follow-up) is more likely to indicate the Subtype III.

Abbreviations: Aβ-42 = the 42 amino acid form of amyloid-β; CSF = cerebrospinal fluid; ESS = Epworth sleepiness score; GDS = Geriatric depression scale; HVLT = Hopkins Verbal Learning Test; LNS = Letter-number sequencing; MCI = mild cognition impairment; MDS-UPDRS = Movement Disorders Society–revised Unified Parkinson's Disease Rating Scale; MoCA = Montreal Cognitive Assessment; MRI-PC = principal component learned from the magnetic resonance imaging; PIGD = postural instability and gait disorder; PPMI = the Parkinson's Progression Markers Initiative; RBD = REM sleep behavior disorder; SCOPA = Scales for Outcomes in Parkinson's Disease; SDM = Symbol–Digit Matching; SHAP = the novel SHapley Additive exPlanation (SHAP) values; STAI = State-Trait Anxiety Inventory.

Tables

Table 1. Demographics and baseline clinical characteristics by subtypes in the PPMI cohort

Variables	Subtype I (Mild progression)	Subtype II (Moderate progression)	Subtype III (Rapid progression)	P-value[a]	Post-hoc[b]	P-value adjusted[c]
# of patients	**145**	**207**	**54**	**-**	**-**	**-**
Age at onset, year	58.1 (9.9)	59.4 (10.0)	64.4 (8.6)	<0.001	III vs. rest	-
Sex male (%)	67 (46.2)	155 (74.9)	44 (81.5)	<0.001	-	-
Race white (%)	140 (96.6)	196 (94.7)	50 (92.6)	0.422	-	-
Symptom duration, month	0.593 (0.682)	0.546 (0.742)	0.667 (0.752)	0.526	-	-
Genetic risk score[d]	-0.009 (0.004)	-0.009 (0.004)	-0.01 (0.004)	0.615	-	0.879
Family history (%)	140 (96.6)	196 (94.7)	50 (92.6)	0.711	-	-
Education history group					-	-
< 12 years	10(6.9)	9(4.4)	7(13.0)	0.102	-	-
12-16 years	90(62.1)	132(63.8)	26(48.2)			
> 16 years	45(31.0)	66(31.9)	21(38.9)			
Motor manifestations						
MDS-UPDRS Part II	5.2 (3.7)	5.7 (4.2)	7.9 (4.4)	<0.001	III vs. rest	<0.001
MDS-UPDRS Part III	19.5 (8.8)	20.7 (8.4)	24.8 (9.6)	<0.001	III vs. rest	0.012
H&Y Stage	1.55 (0.51)	1.54 (0.49)	1.70 (0.49)	0.093	-	0.340
Schwab and England score	93.9 (5.9)	93.2 (5.6)	90.9 (5.7)	0.005	III vs. rest	0.009
Tremor score	0.46 (0.32)	0.50 (0.31)	0.55 (0.34)	0.14	-	0.321
PIGD score	0.22 (0.25)	0.20 (0.19)	0.34 (0.26)	<0.001	III vs. rest	<0.001
Motor phenotype (%)						
Tremor	98 (67.6)	154 (74.4)	33 (61.1)	0.235	-	-
Indeterminate	18 (12.4)	22 (10.6)	6 (11.1)			
PIGD	29 (20.0)	31 (15.0)	15 (27.8)			
Non-motor manifestations						
MDS-UPDRS Part I	5.2 (4.2)	5.4 (3.8)	6.6 (4.4)	0.076	-	0.029
Hallucination	0.01 (0.12)	0.03 (0.18)	0.04 (0.19)	0.467	-	0.51
Apathy	0.18 (0.51)	0.21 (0.48)	0.20 (0.49)	0.82	-	0.838

Pain	0.67 (0.83)	0.72 (0.83)	0.65 (0.84)	0.806	-	0.427
Fatigue	0.56 (0.71)	0.67 (0.79)	0.74 (0.84)	0.272	-	0.092
Sleep						
Epworth sleepiness score	4.9 (3.4)	6.0 (3.1)	6.4 (3.8)	0.001	I vs. rest	0.004
REM sleep behavior disorder	3.4 (2.4)	4.2 (2.7)	5.1 (2.9)	<0.001	1 vs. rest	0.001
QUIP (Impulse control disorders)	0.27 (0.62)	0.31 (0.64)	0.17 (0.46)	0.284	-	0.288
Geriatric depression scale	2.26 (2.79)	2.23 (2.801)	2.87 (2.41)	0.212	-	0.121
State trait anxiety index						
State subscore	32.3 (10.8)	33.2 (9.9)	33.2 (9.5)	0.716	-	0.424
Trait subscore	31.8 (10.3)	32.7 (8.8)	32.4 (9.7)	0.713	-	0.198
SCOPA autonomic questionnaire						
Gastrointestinal (up+down)	1.81 (1.93)	2.1 (1.95)	3.13 (2.35)	<0.001	III vs. rest	<0.001
Urinary	3.88 (2.63)	4.21 (2.89)	5.41 (7.45)	0.040	I vs. III	0.238
Cardiovascular	0.39 (0.68)	0.51 (0.86)	0.52 (0.66)	0.309	-	0.141
Thermoregulatory	0.46 (0.87)	0.47 (0.84)	0.24 (0.72)	0.19	-	0.28
Pupillomotor	0.35 (0.58)	0.42 (0.64)	0.50 (0.79)	0.289	-	0.134
Skin	0.69 (0.88)	0.68 (0.91)	0.83 (0.96)	0.53	-	0.155
Sexual	4.47 (6.48)	3.43 (5.68)	4.48 (6.71)	0.233	-	0.585
Total (sum all)	12.04 (8.39)	11.81 (8.32)	15.11 (11.16)	0.044	II vs. III	0.020
Cognitive function						
MoCA-visuospatial	4.6 (0.8)	4.5 (0.8)	4.4 (0.8)	0.23	-	0.221
MoCA-naming	3.0 (0.2)	2.9 (0.3)	2.9 (0.4)	0.043	I vs. III	0.136
MoCA-attention	5.8 (0.5)	5.8 (0.6)	5.7 (0.7)	0.532	-	0.529
MoCA-language	2.7 (0.6)	2.5 (0.7)	2.6 (0.7)	0.013	I vs. II	0.026
MoCA-delayed recall	3.7 (1.3)	3.3 (1.4)	2.9 (1.6)	<0.001	I vs. rest	0.057
MoCA total score	27.9 (2.0)	27.0 (2.4)	26.7 (2.4)	<0.001	I vs. rest	0.009
Benton judgment of line orientation	13.0 (2.1)	12.9 (2.1)	12.1 (2.2)	0.036	III vs. rest	0.005
HVLT-total recall	25.5 (4.6)	24.3 (5.1)	22.0 (4.7)	<0.001	III vs. rest	0.05
HVLT-delayed recall	8.8 (2.4)	8.3 (2.5)	7.2 (2.5)	<0.001	III vs. rest	0.104
HVLT-discrimination recognition	10.1 (2.3)	9.7 (2.5)	8.5 (2.9)	<0.001	III vs. rest	0.012
HVLT-retention	0.9 (0.2)	0.8 (0.2)	0.8 (0.2)	0.072	-	0.615
LNS	11.4 (2.7)	10.3 (2.4)	9.5 (2.8)	<0.001	I vs. rest	<0.001
Semantic fluency	52.1 (12.1)	48.3 (10.5)	41.3 (11.4)	<0.001	All comparisons	<0.001
Symbol digit test	47.0 (8.6)	44.6 (9.0)	40.7 (9.9)	<0.001	All comparisons	<0.001
CSF biomarkers						
α-synuclein	1607.1 (734.9)	1487.3 (661.3)	1357.9 (505.6)	0.059	-	0.049
Aβ-42	970.1 (463.8)	905.1 (387.7)	781.5 (351.6)	0.020	I vs. III	0.031

P-tau	14.3 (5.4)	14.0 (5.1)	15.0 (6.4)	0.522	-	0.897
T-tau	169.6 (59.4)	163.5 (54.1)	170.1 (66.7)	0.567	-	0.799
Aβ-42/T-tau ratio	5.7 (1.5)	5.6 (1.6)	4.9 (1.9)	0.006	III vs. rest	0.055
Aβ-42/α-synuclein ratio	0.64 (0.22)	0.66 (0.28)	0.59 (0.21)	0.251	-	0.383
P-tau/α-synuclein ratio	0.009 (0.002)	0.010 (0.002)	0.011 (0.003)	<0.001	III vs. rest	<0.001

[a] P-values were calculated using ANOVA (for continuous variables) and χ^2 test (for categorical variables) where appropriate.
[b] Post-hoc analysis was performed using the Tukey HSD test when the ANOVA P-value < 0.05.
[c] ANCOVA was used to calculate p-values (for continuous variables) adjusting for age and sex.
[d] Genetic risk score were calculated based on 90 PD-related loci reported in the latest Genome wide association study[14].

Abbreviations: Aβ-42 = the 42 amino acid form of amyloid-β; CSF = cerebrospinal fluid; HVLT = Hopkins Verbal Learning Test; LNS = Letter-number sequencing; MDS-UPDRS = Movement Disorders Society–revised Unified Parkinson's Disease Rating Scale; MoCA = Montreal Cognitive Assessment; PIGD = postural instability and gait disorder; PPMI = the Parkinson's Progression Markers Initiative; SCOPA = Scales for Outcomes in Parkinson's Disease.

Table 2. Annual progression rates in clinical manifestations and CSF biomarkers by subtypes assessed by linear mixed effects models in the PPMI cohort

Variable	**Subtype I (Mild progression)**		**Subtype II (Moderate progression)**		**Subtype III (Rapid progression)**	
	β	**P value**	**β**	**P value**	**β**	**P value**
Motor manifestations						
UPDRS-Part II	0.42 (0.24, 0.61)	<0.001	1.05 (0.91, 1.19)	<0.001	2.01 (1.55, 2.47)	<0.001
UPDRS-Part III	1.04 (0.64, 1.45)	<0.001	2.10 (1.77, 2.43)	<0.001	3.10 (2.14, 4.06)	<0.001
H&Y Stage	0.07 (0.05, 0.09)	<0.001	0.09 (0.07, 0.10)	<0.001	0.15 (0.10, 0.20)	<0.001
Schwab and England score	-0.97 (-1.25, -0.69)	<0.001	-1.74 (-1.99, -1.49)	<0.001	-3.95 (-5.00, -2.90)	<0.001
Tremor score	-0.00 (-0.02, 0.01)	0.838	0.03 (0.02, 0.04)	<0.001	-0.00 (-0.03, 0.02)	0.864
PIGD score	0.04 (0.02, 0.05)	<0.001	0.06 (0.05, 0.07)	<0.001	0.19 (0.13, 0.24)	<0.001
Non-motor manifestations						
UPDRS-Part I	0.60 (0.45, 0.75)	<0.001	0.91 (0.78, 1.05)	<0.001	1.82 (1.48, 2.16)	<0.001
Hallucination	0.02 (0.01, 0.03)	<0.001	0.03 (0.02, 0.03)	<0.001	0.11 (0.06, 0.16)	<0.001

Apathy	0.01 (-0.01, 0.03)	0.198	0.05 (0.04, 0.07)	<0.001	0.12 (0.06, 0.18)	<0.001
Pain	0.03 (0.01, 0.06)	0.019	0.07 (0.04, 0.09)	<0.001	0.14 (0.10, 0.18)	<0.001
Fatigue	0.05 (0.03, 0.08)	<0.001	0.08 (0.06, 0.11)	<0.001	0.19 (0.14, 0.24)	<0.001
Sleep						
Epworth sleepiness score	0.17 (0.04, 0.30)	0.009	0.48 (0.36, 0.59)	<0.001	1.08 (0.80, 1.35)	<0.001
REM sleep behavior disorder	0.18 (0.09, 0.28)	<0.001	0.26 (0.18, 0.33)	<0.001	0.15 (-0.01, 0.32)	0.074
QUIP (Impulse control disorders)	0.02 (-0.01, 0.04)	0.215	0.06 (0.03, 0.08)	<0.001	0.02 (-0.02, 0.06)	0.319
Geriatric depression scale	-0.11 (-0.20, -0.02)	0.017	0.12 (0.04, 0.19)	0.002	0.48 (0.25, 0.70)	<0.001
State trait anxiety index						
State subscore	-0.54 (-0.84, -0.25)	<0.001	0.04 (-0.21, 0.28)	0.765	0.80 (0.22, 1.37)	0.009
Trait subscore	-0.30 (-0.56, -0.04)	0.027	0.19 (-0.04, 0.42)	0.11	1.09 (0.49, 1.68)	<0.001
SCOPA autonomic questionnaire						
Gastrointestinal	0.27 (0.19, 0.35)	<0.001	0.30 (0.24, 0.36)	<0.001	0.39 (0.24, 0.54)	<0.001
Urinary	0.13 (0.05, 0.21)	0.002	0.28 (0.17, 0.38)	<0.001	0.33 (-0.09, 0.74)	0.131
Cardiovascular	0.05 (0.02, 0.08)	0.003	0.07 (0.04, 0.09)	<0.001	0.20 (0.12, 0.28)	<0.001
Thermoregulatory	0.02 (-0.01, 0.06)	0.251	0.06 (0.03, 0.09)	<0.001	0.07 (0.02, 0.12)	0.015
Pupillomotor	0.02 (-0.00, 0.05)	0.054	0.04 (0.02, 0.06)	<0.001	0.07 (0.02, 0.13)	0.005
Skin	0.03 (0.00, 0.07)	0.047	0.10 (0.07, 0.13)	<0.001	0.14 (0.06, 0.23)	0.001
Sexual	0.14 (-0.08, 0.37)	0.215	0.37 (0.21, 0.53)	<0.001	0.67 (0.29, 1.06)	0.001
Total	0.67 (0.42, 0.93)	<0.001	1.21 (0.97, 1.45)	<0.001	1.85 (1.11, 2.58)	<0.001
Cognitive function						
MoCA	0.07 (-0.01, 0.14)	0.082	-0.07 (-0.15, 0.00)	0.053	-0.98 (-1.31, -0.66)	<0.001
Benton judgment of line orientation	0.02 (-0.04, 0.08)	0.538	-0.06 (-0.11, -0.00)	0.038	-0.28 (-0.42, -0.13)	<0.001
HVLT-total recall	0.23 (0.07, 0.39)	0.006	0.04 (-0.10, 0.17)	0.605	-0.98 (-1.23, -0.72)	<0.001
HVLT-delayed recall	0.11 (0.04, 0.19)	0.004	0.01 (-0.06, 0.08)	0.753	-0.55 (-0.73, -0.36)	<0.001
HVLT-discrimination recognition	0.23 (0.14, 0.32)	<0.001	0.15 (0.08, 0.22)	<0.001	0.04 (-0.15, 0.23)	0.682
HVLT-retention	0.00 (-0.00, 0.01)	0.183	-0.00 (-0.01, 0.00)	0.582	-0.05 (-0.07, -0.03)	<0.001
LNS	-0.09 (-0.16, -0.01)	0.029	-0.10 (-0.17, -0.04)	0.001	-0.47 (-0.67, -0.27)	<0.001
Semantic fluency	0.26 (-0.10, 0.62)	0.166	-0.28 (-0.51, -0.04)	0.021	-1.49 (-2.11, -0.87)	<0.001
Symbol digit test	0.35 (-0.02, 0.73)	0.067	-0.14 (-0.39, 0.10)	0.252	-1.26 (-1.93, -0.59)	<0.001
CSF biomarkers						
α-synuclein	-22.82 (-47.31, 1.67)	0.07	-10.30 (-22.89, 2.29)	0.111	13.57 (-13.06, 40.19)	0.325
Aβ-42	-6.33 (-17.93, 5.26)	0.287	-10.33 (-18.17, -2.49)	0.011	-20.99 (-37.82, -4.17)	0.022
P-tau	0.06 (-0.03, 0.15)	0.219	0.06 (-0.02, 0.15)	0.162	0.14 (-0.09, 0.36)	0.239
T-tau	0.62 (-0.50, 1.75)	0.279	1.02 (-0.15, 2.19)	0.089	2.10 (-0.32, 4.51)	0.102
Aβ-42/T-tau ratio	-0.04 (-0.10, 0.01)	0.146	-0.07 (-0.11, -0.03)	<0.001	-0.16 (-0.25, -0.08)	<0.001
Aβ-42/α-synuclein ratio	0.00 (-0.01, 0.01)	0.888	-0.00 (-0.01, 0.00)	0.294	-0.01 (-0.02, -0.00)	0.031

P-tau/α-synuclein ratio	0.00 (-0.00, 0.00)	0.108	0.00 (-0.00, 0.00)	0.265	0.00 (-0.00, 0.00)	0.552

Abbreviations: Aβ-42 = the 42 amino acid form of amyloid-β; CSF = cerebrospinal fluid; HVLT = Hopkins Verbal Learning Test; LNS = Letter-number sequencing; MDS-UPDRS = Movement Disorders Society–revised Unified Parkinson's Disease Rating Scale; MoCA = Montreal Cognitive Assessment; PIGD = postural instability and gait disorder; PPMI = the Parkinson's Progression Markers Initiative; SCOPA = Scales for Outcomes in Parkinson's Disease.