

ARTICLE OPEN



Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts

Anant Dadu^{1,2,3}, Vipul Satone⁴, Rachneet Kaur⁴, Sayed Hadi Hashemi¹, Hampton Leonard^{2,3,5}, Hirotaka Iwaki^{2,3,5}, Mary B. Makarious^{5,6,7}, Kimberley J. Billingsley⁵, Sara Bandres-Ciga^{2,5}, Lana J. Sargent^{2,8}, Alastair J. Noyce^{7,9}, Ali Daneshmand¹⁰, Cornelis Blauwendraat^{10,2,5}, Ken Marek^{11,12}, Sonja W. Scholz^{13,14}, Andrew B. Singleton^{10,2,5}, Mike A. Nalls^{2,3,5}, Roy H. Campbell¹⁰ and Faraz Faghri^{10,2,3,5}✉

The clinical manifestations of Parkinson's disease (PD) are characterized by heterogeneity in age at onset, disease duration, rate of progression, and the constellation of motor versus non-motor features. There is an unmet need for the characterization of distinct disease subtypes as well as improved, individualized predictions of the disease course. We used unsupervised and supervised machine learning methods on comprehensive, longitudinal clinical data from the Parkinson's Disease Progression Marker Initiative ($n = 294$ cases) to identify patient subtypes and to predict disease progression. The resulting models were validated in an independent, clinically well-characterized cohort from the Parkinson's Disease Biomarker Program ($n = 263$ cases). Our analysis distinguished three distinct disease subtypes with highly predictable progression rates, corresponding to slow, moderate, and fast disease progression. We achieved highly accurate projections of disease progression 5 years after initial diagnosis with an average area under the curve (AUC) of 0.92 (95% CI: 0.95 ± 0.01) for the slower progressing group (PDvec1), 0.87 ± 0.03 for moderate progressors, and 0.95 ± 0.02 for the fast-progressing group (PDvec3). We identified serum neurofilament light as a significant indicator of fast disease progression among other key biomarkers of interest. We replicated these findings in an independent cohort, released the analytical code, and developed models in an open science manner. Our data-driven study provides insights to deconstruct PD heterogeneity. This approach could have immediate implications for clinical trials by improving the detection of significant clinical outcomes. We anticipate that machine learning models will improve patient counseling, clinical trial design, and ultimately individualized patient care.

npj Parkinson's Disease (2022)8:172; <https://doi.org/10.1038/s41531-022-00439-z>

INTRODUCTION

Parkinson's disease (PD) is a complex, age-related neurodegenerative disease that is defined by a combination of core diagnostic features, including bradykinesia, rigidity, tremor, and postural instability^{1,2}. Substantial phenotypic heterogeneity is well recognized within the disease, complicating the design and interpretation of clinical trials, and limiting patients' counseling about their prognosis. The clinical manifestations of PD vary by age at onset, rate of progression, associated treatment complications, as well as the occurrence and constellation of motor/nonmotor features.

The phenotypic heterogeneity that exists within the PD population poses a major challenge for clinical care and clinical trial design. A clinical trial has to be suitably powered to account for interindividual variability, and as a consequence, trials are either large, long, expensive, and/or only powered to see large effects. This problem becomes particularly burdensome as we move increasingly towards early stage trials when therapeutic interventions are likely to be most effective. To that effect, defining subcategories of PD and the ability to predict even a proportion of the disease course has the potential to significantly

improve cohort selection, inform clinical trial design, reduce the cost of clinical trials, and increase the ability of such trials to detect treatment effects.

Attempts thus far at the characterization of disease subtypes have followed a path of clinical observation based on age at onset or categorization based on the most observable features³. Thus, the disease is often separated into early-onset versus late-onset disease, slowly-progressing "benign" versus fast-progressing "malignant" subtypes, PD with or without dementia, or based on the most prominent clinical signs into a tremor-dominant versus a postural instability with gait disorder subtype^{4,5}. This dichotomous separation, while intuitive, does not faithfully represent the clinical features of the disease, which are quantitative, complex, and interrelated. A more realistic representation of the disease and disease course requires a transition to a data-driven, multi-dimensional schema that encapsulates the constellation of interrelated features and allows tracking (and ultimately predicting) change^{6,7}.

Previous studies used cluster analysis, a data-driven approach, to define two to three clinical PD subtypes^{8–13}. Depth of

¹Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA. ²Center for Alzheimer's and Related Dementias (CARD), National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA. ³Data Tecnica International, Washington, DC 20812, USA. ⁴Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA. ⁵Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA. ⁶Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, London, UK. ⁷UCL Movement Disorders Centre, University College London, London, UK. ⁸School of Nursing, Virginia Commonwealth University, Richmond, VA 23298, USA. ⁹Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London and Department of Neurology, Royal London Hospital, London, UK. ¹⁰Department of Neurology, Boston Medical Center, Boston University School of Medicine, Boston, MA 02118, USA. ¹¹InviCRO LLC, Boston, MA, USA. ¹²Molecular Neuroimaging, A Division of InviCRO, New Haven, CT, USA. ¹³Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA. ¹⁴Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ✉email: faraz.faghri@nih.gov

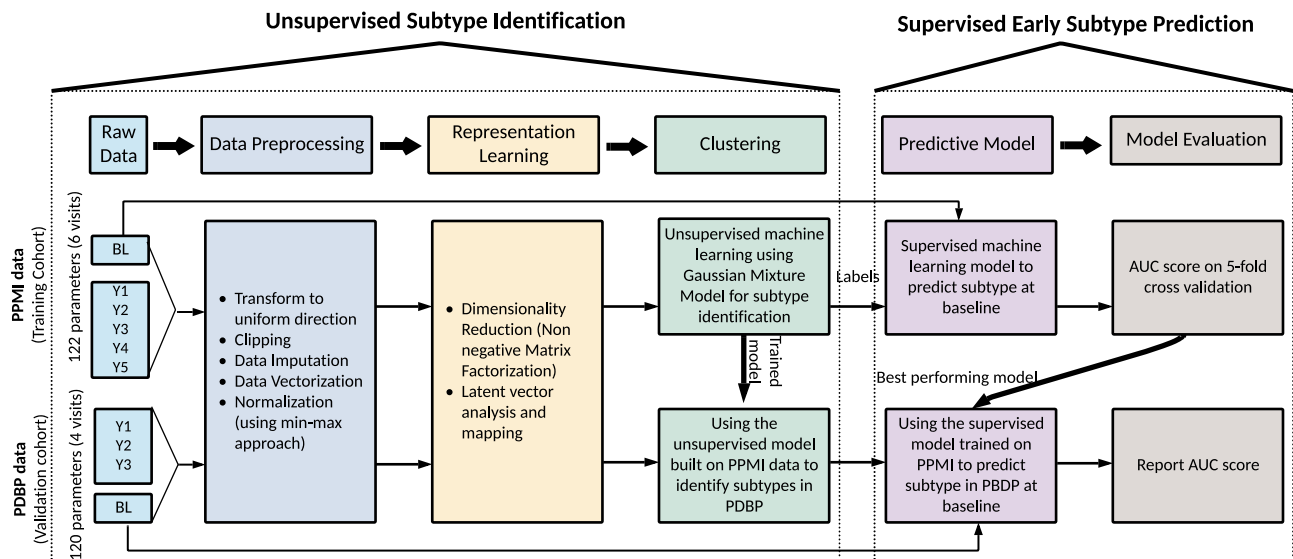


Fig. 1 Workflow of analysis and model development. PPMI Parkinson's Progression Marker Initiative, PDBP Parkinson's Disease Biomarkers Program, BL Baseline, Y1 Year1, Y2 Year2, Y3 Year3, Y4 Year4, Y5 Year5, AUC Area under receiver operating characteristic curve.

phenotypic information and longitudinal assessments in these studies were variable and often limited to certain clinical features and short-term follow-ups. Moreover, many previous studies were limited by insufficient methods to capture longitudinal changes over multiple assessment visits. To this date, none of the previous approaches to PD clustering were replicated in an independent cohort with transparent code and analysis.

We have previously used multi-modal data to produce a highly accurate disease status classification and to distinguish PD-mimic syndromes from PD¹⁴. These efforts demonstrated the utility of data-driven approaches in the dissection of complex traits and have also led us to the next logical step in disease prediction: supplementing the prediction of whether a person has or will have PD also to include a prediction of the timing and directionality of the course of their disease.

Here, we describe our work on delineating and predicting the clinical progression of PD and for a workflow of the analysis, please refer to Fig. 1. The first stage of this effort requires creating a multi-dimensional space that captures the disease's features and the progression rate of these features (i.e., velocity). Rather than creating a space based on a priori concepts of differential symptoms, we used data dimensionality reduction methods on the complex clinical features observed 60 months after initial diagnosis to create a meaningful spatial representation of each patient's status at this time point. After creating this space, we used unsupervised clustering to determine whether there were clear subtypes of disease within this space. This effort identified three distinct clinical subtypes corresponding to three groups of patients progressing at varying velocities (i.e., slow, moderate, and fast progressors). These subtypes were validated and replicated in an independent cohort. Following the successful creation of disease subtypes within a progression space, we created a baseline predictor that accurately predicted an individual patient's clinical group membership 5 years later. Further, we examined the predictive capability of biospecimen biomarkers at baseline and the genetic information in identifying the subtypes. Our work highlights the utility of machine learning as an ancillary diagnostic tool to identify disease subtypes and project individualized progression rates based on model predictions.

RESULTS

Clustering vectors of progression

Figure 2 shows the result of the mathematical projection of PD progression, called *Parkinson's disease progression space* detailing normalized progression trajectories of each sample relative to others based on this unsupervised classification system. This space shows the relative progression velocity of each patient in 60 months (i.e., speed and direction). The progression velocity level is divided into three main dimensions: motor, cognitive, and sleep-related disturbances. Movement disorders specialists audited component features to categorize these clinical measures into domains of sleep, motor, and cognition disturbance after identification by the algorithm. Based on latent variables clustered within the Parkinson's progression space, the projected motor dimension was responsible for 63.58% of the explained variance, followed by the sleep dimension (21.81%), and cognitive dimension (14.61%). Motor symptoms are the hallmark of PD progression, with sleep and cognitive decline being, in some cases, elevated past that decline seen in controlled aging. The projected motor dimension significantly contributes towards PD progression; however, sleep and cognition are essential, accounting for 37% variation. Across these trajectories, the unsupervised learning analysis reveals and classifies patients into three main subtypes of PD, relating to rates of disease progression: slow progressors (PDvec1), moderate progressors (PDvec2), and fast progressors (PDvec3). This shows how we can map the clinical features and progression velocity from the point of diagnosis. The components of the motor, cognitive, and sleep dimensions with a description of the latent space used to define the progression space that may aid in interpretability are shown in Supplementary Fig. 1 (see Supplementary Material for details).

Identified Subtypes and their Clinical Characteristics

Figure 3 shows the visualization of unsupervised learning via Gaussian Mixture Model (GMM) in a two-dimensional progression space. In two-dimensional progression space, the projected dimensions represent motor (y-axis) and cognitive (combined with sleep) (x-axis) components. Projected dimensions are normalized; the increase in values along either direction signifies a higher decline. GMM fits the data into different subtypes relating to velocity of decline across symptomatology from non-PD

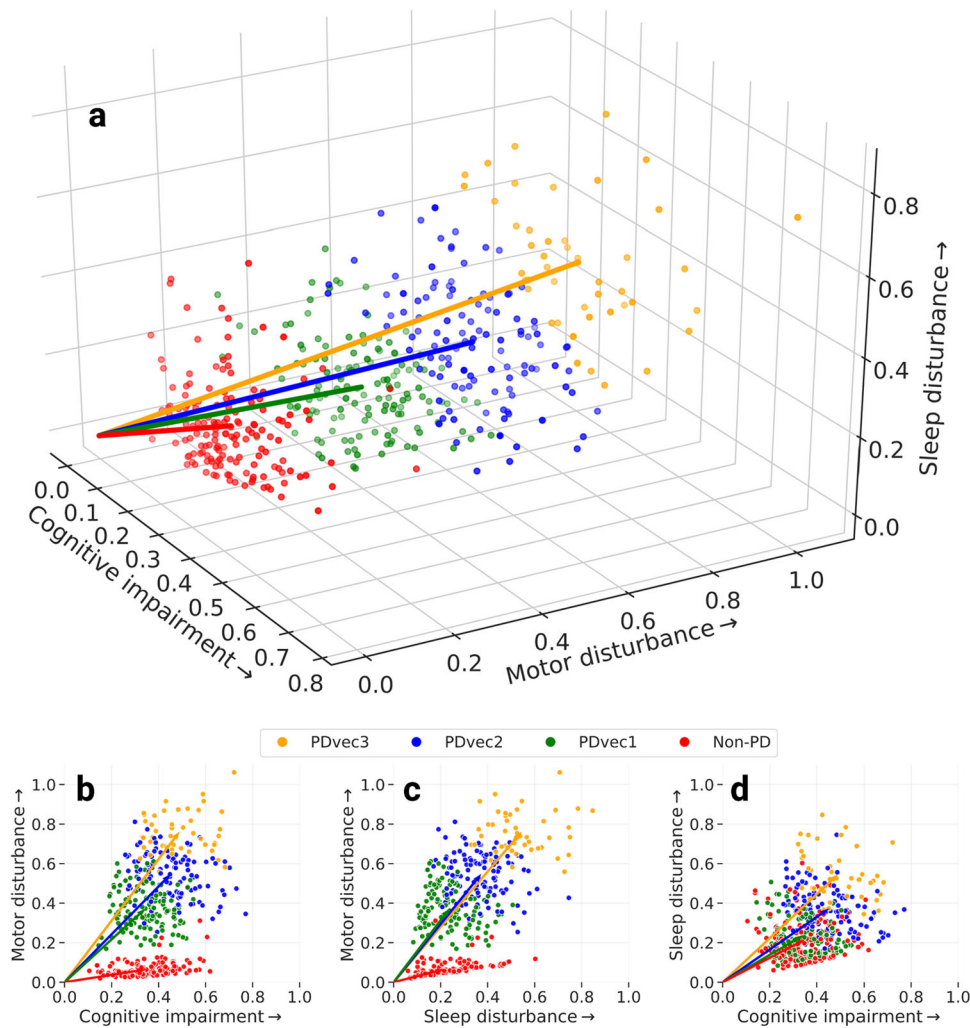


Fig. 2 Different views of the Parkinson's disease progression space in 5 years with three corresponding projected dimensions (cognitive, motor, and sleep dimensions) on a normalized scale. Subtypes of PD are identified using unsupervised learning (PDvec1, PDvec2, and PDvec3). **a** Shows the view of all three dimensions, **b** view of the motor and cognitive dimensions, **c** view of motor and sleep dimensions, and **d** view of sleep and motor dimensions.

controls. The Bayesian information criterion has identified three Gaussian distributions representing three PD subtypes. Further, mean values of PD subtypes in lower dimensional progression space are significantly different along both Motor dimension (PDvec1 = 0.43 [95%CI: 0.41–0.44], PDvec2 = 0.64 [95%CI: 0.62–0.65], PDvec3 = 0.89 [95%CI: 0.85–0.92]) and Cognitive/Sleep dimensions (PDvec1 = 0.40 [95%CI: 0.39–0.42], PDvec2 = 0.57 [95%CI: 0.56–0.59], PDvec3 = 0.71 [95%CI: 0.68–0.75]) all with non-overlapping CIs across groups. These three groups identified algorithmically within the case population change over time differently within the progression space and across specific biomarkers of progression, with PDvec3 generally progressing at a much steeper slope (Supplementary Figs. 2, 3, 4, 5). Details of this can be seen in the Supplementary Material section describing the Five-Year PD Progression Space. Using our proposed approach, 45% (134/294) of PD patients identified as PDvec1 (slow progressors), with 39% (114/294) belonging to PDvec2 (medium progressors) and PDvec3 (fast progressors) accounts for 16% (46/294) patients.

Biological characteristics of the identified subtypes

Figure 4 shows the variation of biological biomarkers for each PD subtype over time. In terms of patients' features, height and

weight show a significant decline over time for the fast progressors (PDvec3) compared to other subtypes. We used a linear mixed effects model for association testing of PD subtypes and serum neurofilament light (Nfl) measurements. PDvec3 has a significantly steeper slope across time than PDvec1 after adjusting for sex, height, weight, and age at baseline ($P < 0.005$). More details are described in the Association testing of Nfl with PD subtypes section of Supplementary Material and Supplementary Table 1. PD patients have lower values compared to healthy controls for all CSF sample measurements such as alpha-synuclein, total tau protein, A β 42, and p-Tau181.

Genetic analysis of the identified subtypes

In terms of the genetic association of PDs identified subtypes, genetic risk scores (GRS) were calculated^{15–17}. As a one-time measurement, the GRS was not included during the longitudinal clustering exercise; however, we analyzed regressions comparing associations between the GRS and either the continuous predicted cluster membership probability (linear regression) or the binary membership in a particular cluster group compared to the others. All models were adjusted for age at onset, biological sex, and principal components as covariates to adjust for population substructure in PPMI. The GRS was significantly associated with

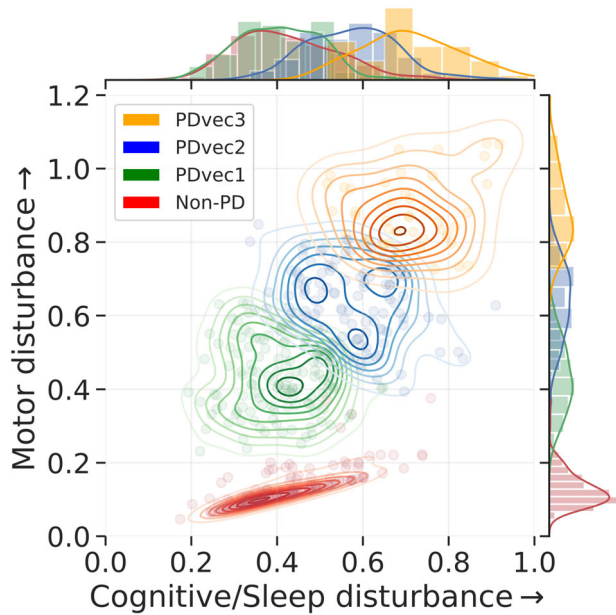


Fig. 3 PD five-year progression space. Visualization of unsupervised learning via GMM on two-dimensional progression space and identification of three Gaussian distributions representing three distinct PD subtypes. An increase in value along either direction reflects the increase in the disturbance on a normalized scale.

decreasing magnitude of the sleep vector per Standard deviation (SD) of increase in the GRS ($\beta = -0.029$, $se = 0.010$, $p = 0.002$, adjusted $r^2 = 0.046$). For binary models of membership, we see that the GRS is weakly but significantly associated with a decreased risk of membership in PDvec3 (odds ratio = 0.563 per 1 SD increase from case GRS mean, $\beta = -0.574$, $se = 0.244$, $P = 0.018$) and increased risk of membership in PDvec1 (odds ratio = 1.341, $\beta = 0.293$, $se = 0.134$, $P = 0.0282$) all relative to the moderate progressing group as a reference. The lack of a strong genetic association is due to the small sample size, and that genetic variants relating to risk do not necessarily affect progression.

Replication in an independent cohort

In order to ensure the generalizability and validity of the results, we replicated the subtype identification in an independent PDBP cohort. Details on differences between the training (PPMI) and replication (PDBP) cohorts can be found in the appropriately named Supplementary Material section. In the PDBP cohort, 46% (121/263) of PD patients were identified as PDvec1, 23% (60/263) belonged to PDvec2, and the remaining 31% (82/263) were classified into the PDvec3 group. We observed less manifested separation of PDvec2 (medium progressors) in the PDBP cohort. In the progression space, the spatial differences between subtypes become more apparent with increased longitudinal data. The reason can be attributed to the fact that the PDBP cohort has 3 years of longitudinal data compared to 5-year data in the PPMI cohort.

Figure 5 shows the identified subtypes in the independent PDBP cohort using the model developed on the PPMI dataset. We see that the identified subtypes in the PDBP cohort are similar to the ones in the PPMI dataset in terms of progression. Due to the limited length of the PDBP study (36 months), the visualization of progression space is shown through the 36 months follow-up from the baseline. The PPMI and PDBP cohorts are clinically different cohorts and recruited from different populations. The replication of our results in the PDBP cohort that was recruited with a different protocol shows the strength of our study's

methodology. We demonstrate that if we ascertain the same phenotypes using standardized scales, we can reliably discern the same subtypes and progression rates. This suggests that our results may be generalizable and the clinical subtypes reproducible.

Supervised early subtype prediction

Following the data-driven organization of subjects into progression subtypes and clustering them into three subtypes, we developed three models to predict patient progression class after 60 months based on varying input factors: (a) from baseline clinical factors, (b) from baseline and year one clinical factor, (c) biological and genetics measurements. Figure 6a and Fig. 6b show the ROC (Receiver Operating Characteristic) curves of our multi-class supervised learning predictors. We correctly distinguish patients with PD based on baseline only input factors and predict their 60-month prognosis with an average AUC of 0.92 (95% CI: 0.94 ± 0.01 for PDvec1, 0.86 ± 0.01 for PDvec2, and 0.95 ± 0.02 for PDvec3) at cross-validation. The predictor built on baseline and year 1 data performs even better with an average AUC of 0.953 (95% CI: 0.97 ± 0.01 for PDvec1, 0.91 ± 0.02 for PDvec2, and 0.97 ± 0.01 for PDvec3) also at cross-validation. In Fig. 6d, we have shown the PD subtype predictive performance at baseline, only using baseline data, and years after, as more data becomes available and combined with the baseline. The increased accuracy trend is due to the availability of more information about a subject. This approach is also practical in a clinical setting, as physicians will provide a better prognosis for patients after a one-year follow-up. Out of identified top-20 features, 11 belong to the motor dimension, 5 are from the sleep dimension and 4 are a part of cognitive dimension, which is in line with the amount of variability explained by each dimension (Supplementary Tables 2, 3, 4). Further details on feature importance contributing to the accuracy of these models can be found in the Supplementary Material section entitled Feature Importance and Supplementary Figs. 6, 7, 8.

Besides the cross-validation of predictive models in the PPMI cohort, we have also validated the accuracy of the predictive model in the independent PDBP cohort. The predictive model trained on the PPMI baseline data correctly distinguished PDBP patients with an AUC of 0.84 (ROC curves in Fig. 6c). The replicated predictive model performs very well for PDvec1 and PDvec3 (AUC of 0.91 and 0.88, respectively). However, due to the small sample size, the predictive model does not predict as well on PDvec2 (AUC of 0.73). Fewer samples make up the PDvec2 cluster in the replication cohort, and it has been easier for the predictive model to predict the more extreme subtypes (i.e., PDvec1 and PDvec3). Despite the smaller sample size of the PDBP cohort, the results strongly validate our previous observations of distinct, computationally discernible subtypes within the PD population. This finding indicates that our methodology is robust, and our unique progression analysis and clustering approach result in the same clusters. In summary, we have mined data to identify three clinically related constellations of symptoms naturally occurring within our longitudinal data that summarize PD progression (63.58%, 21.81%, 14.61% variance loadings) comprised of factors relating to motor, sleep, and cognitive.

Biomarker based prediction

The performance of PD progression prediction models using biomarkers and genetic measurements for the PPMI cohort is shown in Fig. 7 and Fig. 8. Additional notes on model interpretation as well as how the models deal with participants in the study that have changed diagnoses over time can be found in the Supplementary Material in the appropriately named sections and Supplementary Fig. 9.

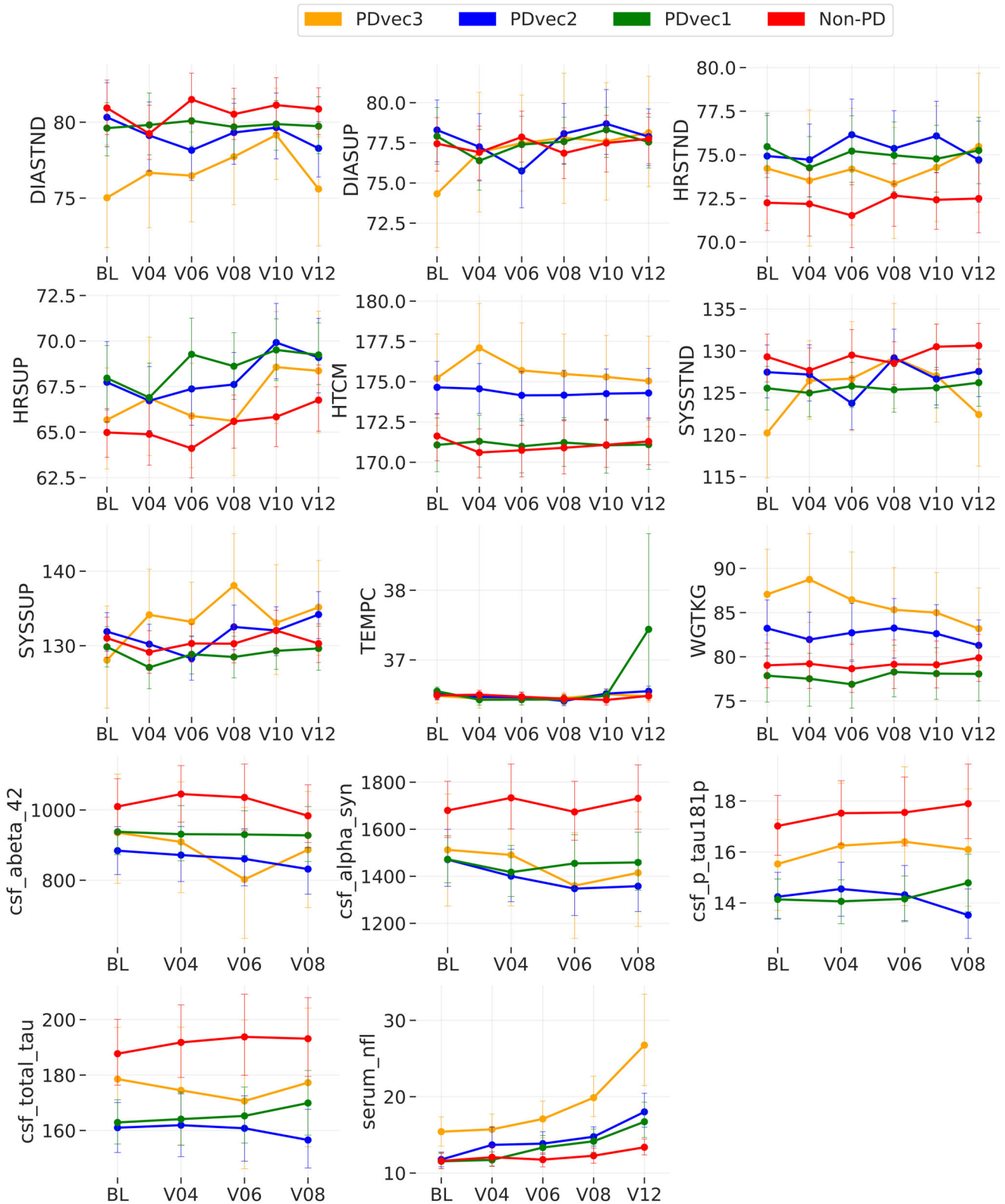


Fig. 4 Shows the biological biomarker variation of each PD subtype over time. The graphs demonstrate the actual clinical values of each subtype overtime for vital signs (DIASSTND standing diastolic blood pressure (BP), DIASUP supine diastolic BP, HRSTND standing heart rate, HRSUP supine heart rate, SYSSTND standing systolic BP, SYSSUP supine systolic BP, HTCM height in cm, TEMPC: temperature in C, WGTKG weight in kg), cerebrospinal fluid (abeta_42 β -amyloid 1-42, alpha_syn alpha-synuclein, p_tau181p phospho-tau181, total_tau total tau protein), and serum neurofilament light levels (serum_nfl). BL: Baseline. V04 visit number 4 after 12 months. V06: visit number 6 after 24 months. V08 visit number 8 after 36 months. V10 visit number 10 after 48 months. V12 visit number 12 after 60 months. In all panels, data is presented as mean \pm s.e.m.

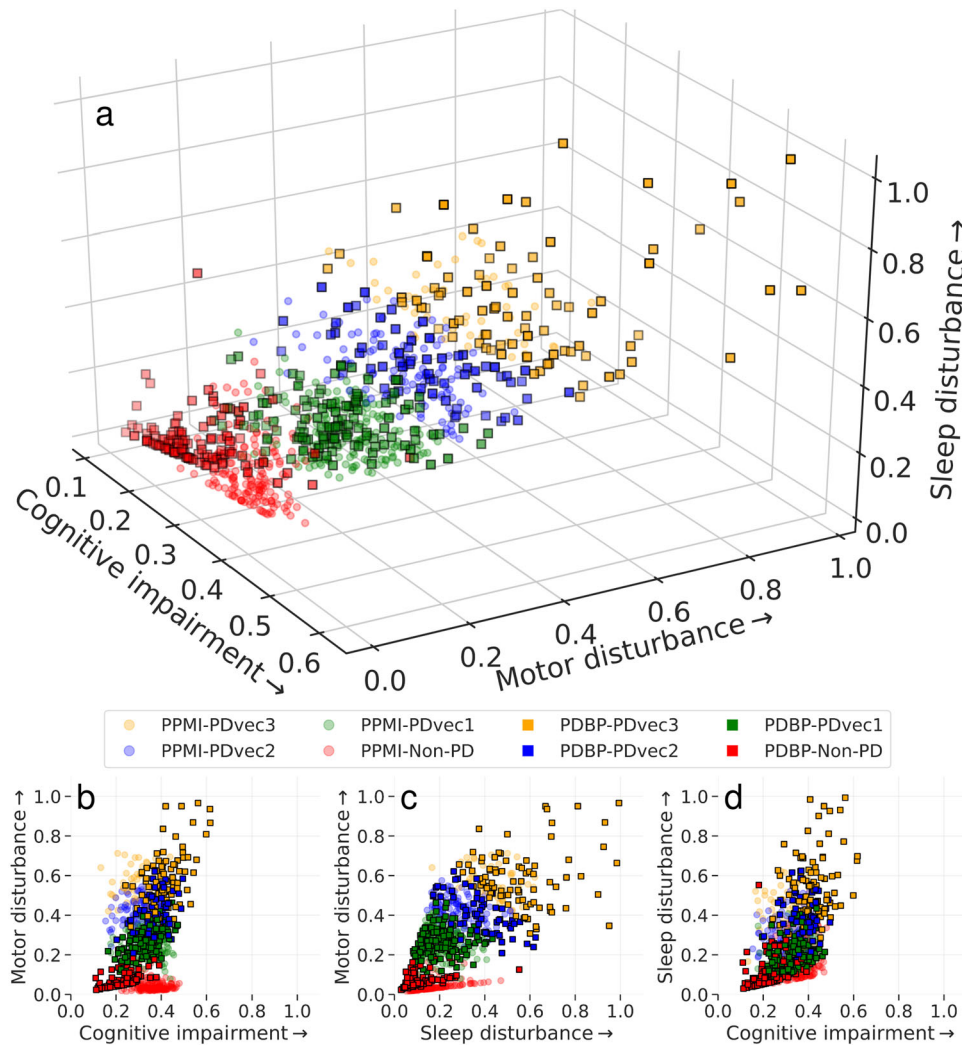


Fig. 5 Shows the identified subtypes in the independent PDBP cohort using the model developed on the PPMI dataset. Similar PDBP and PPMI subtypes in terms of progression. **a** Shows the view of all three dimensions, **b** view of the motor and cognitive dimensions, **c** view of motor and sleep dimensions, and **d** view of sleep and cognitive dimensions. The normalized progression space is shown through the 36 months follow up from baseline for both PPMI and PDBP datasets.

A trained machine learning model using only UPDRS can predict subtypes with a 0.77 AUC score (Fig. 7b) compared to 0.92 with the model that uses all symptomatic clinical measures at baseline on five-fold cross-validation. It demonstrates the utility of machine learning models in integrating features from multiple dimensions to provide an optimal classification performance. Biomarkers, such as age, height, weight, and CSF measurements, are shown to be essential features in predicting the subtypes at baseline (shown in Fig. 8). The mean AUC score is 0.67 (Fig. 7e) using biospecimen, vital signs, and demographics. In comparison, genetic features show slightly lower performance (AUC score 0.66, Fig. 7d). A combination of demographics, biospecimen, vital signs, genetics, and UPDRS is the best performing model (AUC score 0.80, Fig. 7a). It is important to note that segregating PDvec3 (fast progressive subtype) has shown similar performance with only the UPDRS model and that model that includes other biomarkers and genetic measurement. It might be valuable to evaluate the UPDRS model's performance in a clinical setting, as UPDRS is a standard measure of PD diagnosis and disease severity. Further, the individual components (clinical questionnaire responses) of UPDRS are crucial, and machine learning models exploit and utilize their complex interaction to form a composite score of PD subtypes prediction. A simple aggregation (average) of UPDRS

individual responses might not have similar subtype prediction power. Based on the ease of availability in real-world clinical settings, we suggest combining UPDRS, genetics, biomarkers, and demographics as a subtype diagnostic model, which has a 0.80 AUC score (Fig. 7a). Further validation of the model is necessary to improve generalizability in other cohorts to make it an application for clinicians.

DISCUSSION

Prediction of disease and disease course is a critical challenge in the patient counseling, care, treatment, and research of complex heterogeneous diseases. Within PD, meeting this challenge would allow appropriate planning for patients and symptom-specific care (for example mitigating the chance of falls, identifying patients at high risk for cognitive decline or rapid progression, etc.). Perhaps even more importantly at this time, prediction tools would facilitate more efficient execution of clinical trials. With models predicting a patient-specific disease course, clinical trials could be shorter, smaller, and would be more likely to detect smaller effects, thus, decreasing the cost of phase 3 trials dramatically and essentially reducing the exposure of pharmaceutical companies to a typically expensive and failure-prone area.

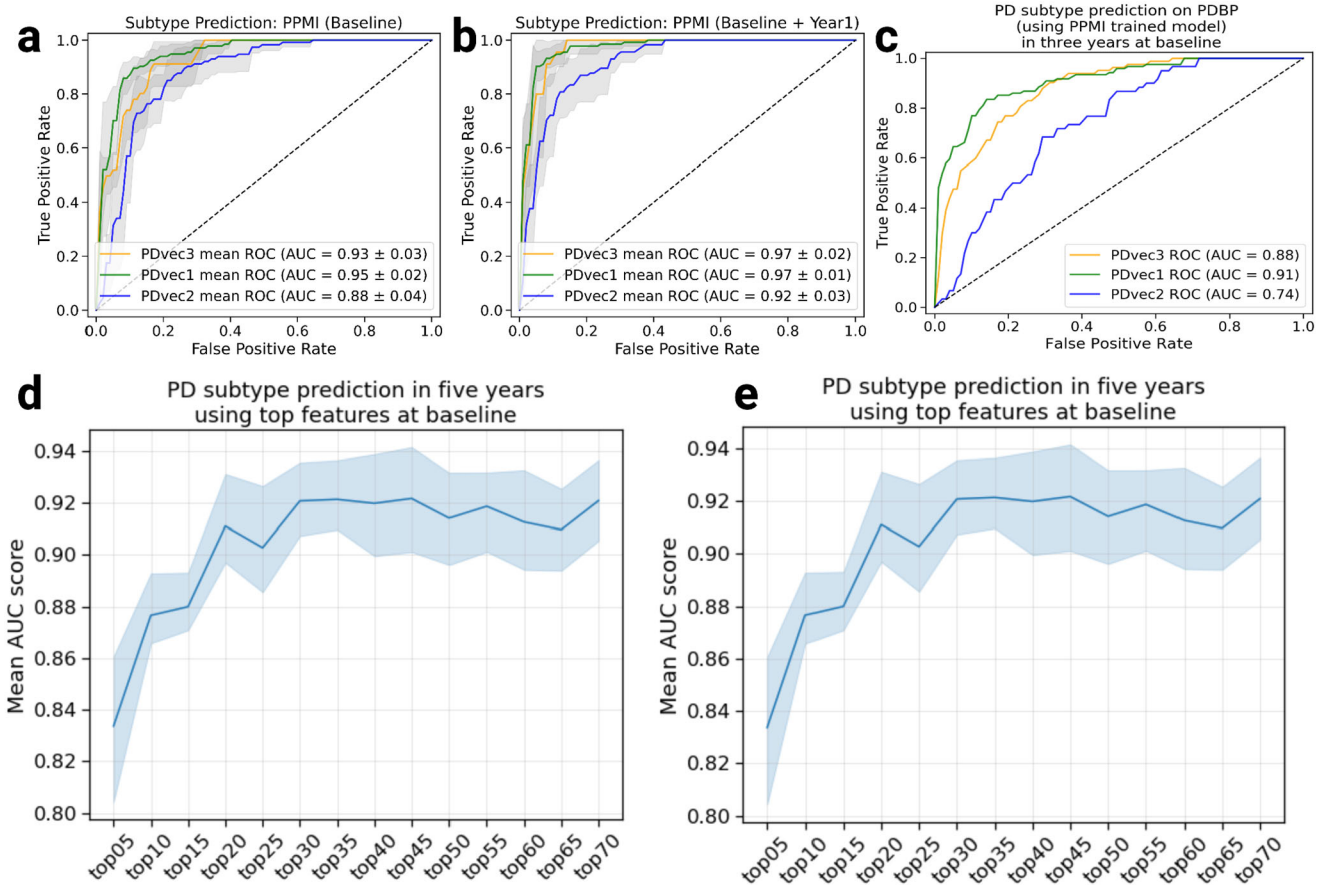


Fig. 6 Shows the performance of Parkinson's disease progression prediction models. **a** The ROC (receiver operating characteristic) for the predictive model at baseline developed on the PPMI cohort evaluated using five-fold cross-validation. **b** The ROC for the predictive model developed on the baseline, and first-year data of the PPMI cohort evaluated using five-fold cross-validation. **c** The ROC for the predictive model developed on the PPMI baseline and tested on the PDBP cohort. **d** Performance of predictive models using data starting from baseline, only using baseline data, and years after, as more data becomes available and combined with the baseline. The y-axis shows the average AUC score across PD subtypes in the PPMI dataset. **e** Contribution of important features to achieve high accuracy. By including only 20 features, we can achieve an AUC of greater than 0.90. In all panels, data is presented as mean \pm s.e.m.

We previously had considerable success in constructing, validating, and replicating a model that allows a data-driven diagnosis of PD and the differentiation of PD-mimic disorders, such as those patients who have parkinsonism without evidence of dopaminergic dysfunction¹⁴. We set out to expand this work by attempting to use a novel approach to (1) define natural subtypes of the disease, (2) attempt to predict these subtypes at baseline, and (3) identify progression rates within each subtype and project progression velocity.

While the work here represents a step forward in our efforts to sub-categorize and predict PD, much more needs to be done. The application of data-driven efforts to complex problems such as this is encouraging; however, the primary limitation of such approaches is that they require large datasets to facilitate model construction, validation, and replication. These datasets should include standardized phenotype collection and recording to achieve the most powerful predictions. Longer follow-ups, more ancestral diversity in samples, and large sample series are crucial to broadening the applicability of this work. Collecting such data is a challenge in PD, with relatively few cohorts available with deep, wide, well-curated data. Thus, a critical need is the expansion or replication of efforts such as PPMI or PDBP, importantly with a model that allows unfettered access to the associated data; the cost associated with this type of data collection is large, but these are an essential resource in our efforts in PD research. Global

Parkinson's Genetics Program (GP2) project has the potential to address some of these limitations in the future (<https://gp2.org/>).

A study used cluster analysis to identify patient subtypes and their corresponding progression rates¹⁰, although these used percentile cutoffs and are not completely data-driven in nature. However, this study evaluated clusters according to only two-time points, baseline, and short-term follow-up, that were aggregated into a Global Composite Outcome score. In return, the subtypes did not capture the fluctuations in the prognosis of subtypes. More recently, a study used a Long-Short Term Memory-based deep learning algorithm to discover PD subtypes, with each subtype showing different progression rate¹⁸. The loss of interpretability with deep learning models makes their approach less suitable for practical purposes. Another study proposed a trajectory-based clustering algorithm to create patient clusters based on trajectory similarity¹⁹. The algorithm gives equal importance to all the features; however, PD is a multi-dimensional spectrum of symptoms with overlapping features derived from simultaneous pathological processes²⁰. Finally, in order to be used in practice, subtyping solutions need to be replicated in a different cohort to show the reliability of methods in assigning individual patients to a subtype. Additionally, none of these previous studies used completely independent replication data.

Our findings can also have implications for the day-to-day practice of clinicians. Movement disorders specialists often use

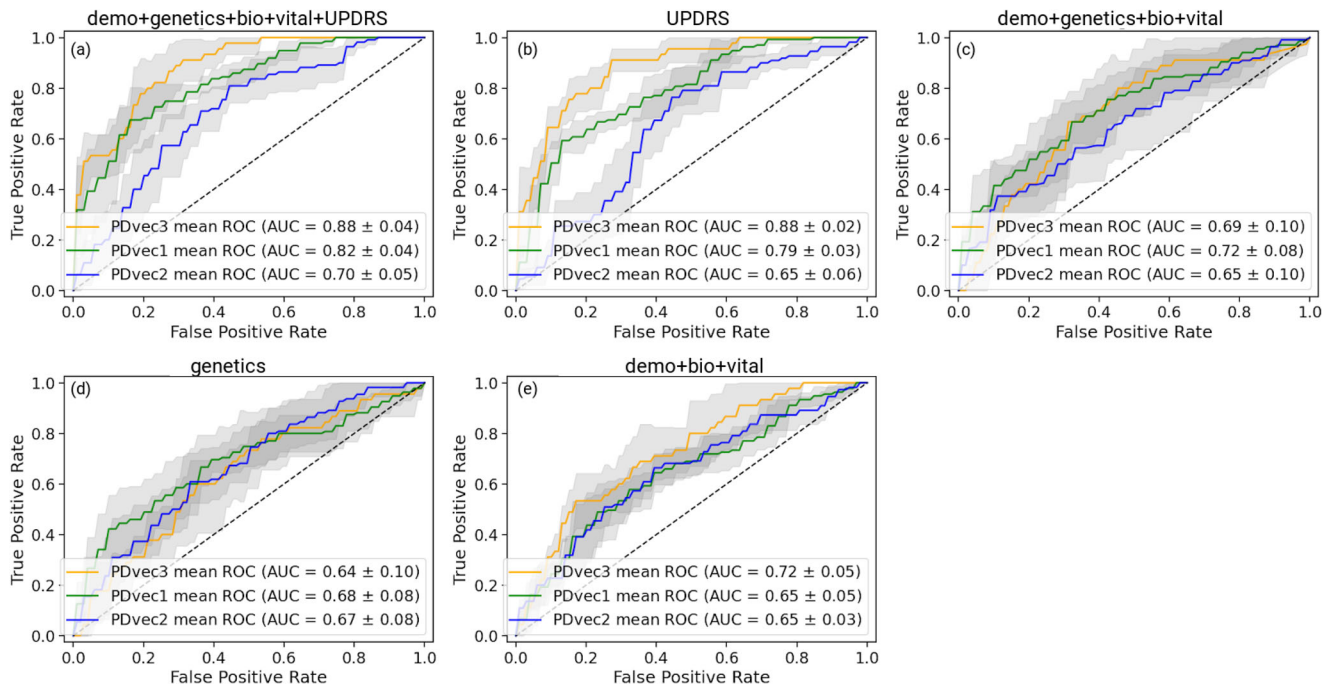


Fig. 7 Shows the performance of Parkinson's disease progression prediction models using biomarkers and genetic measurements for the PPMI cohort. All models are evaluated using five-fold cross-validation. From top left to bottom right: **a** The ROC for the predictive model using a combination of demographics (education, year, sex, race), biospecimen (cerebrospinal fluid, serum Nfl levels), genetics (hg genotype), vital signs (weight, height, blood pressure) and UPDRS measurements. **b** The ROC for the predictive model developed on UPDRS scores. **c** The ROC for the predictive model developed using demographics, genetics, vital signs, and biospecimen measurements. **d** The ROC for the predictive model developed on genetic measurements. **e** The ROC for the predictive model uses only demographics, vital signs, and biospecimen measurements. In all panels, data is presented as mean \pm s.e.m.

screening tools such as MDS-UPDRS to assess a patient's progression and response to treatment. However, performing these clinical assessments requires experience, expertise, and time, which hinders its widespread use by other clinicians (and even neurologists who are not trained in movement disorders). Underutilization of clinical assessment tools can lead to the suboptimal characterization of PD patients and their clinical course, which in turn impacts their care. Our study is one of the first of its kind which systematically assessed the accuracy of each feature of MDS-UPDRS in predicting PD's course. For example, daytime sleepiness (NP1SLPD) was found to have the highest importance in clinical progression, followed by doing hobbies and activities (NP2HOB), dressing (NP2DRES), and urinary problems (NP1URIN). Knowing the clinical features with the highest yield in course prediction can help clinicians to tailor their assessment and better inform patients about their disease course. In addition, shortened versions of comprehensive assessment tools can be utilized to address specific clinical questions. Surprisingly, none of the genetic markers studied had high accuracy in clinical course prediction. Our work initiates multiple questions that are worth exploring in the future. The progression space seems to stabilize after 3 years from the baseline. It will be interesting to predict how much time (from baseline) is required to provide reliable predictions about the PD subtypes. Secondly, fast progressors do not worsen with the multiple symptoms such as Epworth and MDS-UPDRS scores (Supplementary Fig. 5) from the fourth to the 5th year, while other subtypes do. It raises the question of whether the fast progressors reach the saturation point after some time from baseline. It will be useful if we can look for similar patterns in other PD datasets. Incorporating imaging data for PD subtypes is also an exciting direction to pursue in the future. Finally, dramatic increases in Nfl and high baseline levels of Nfl could be an indicator of potential rapid progression.

In this study, we addressed the complexities of PD. We integrated unlabeled, multimodal, and longitudinal data. The longitudinal data had a long-term nature, and we were interested in capturing the overall pattern of the individual's trajectories. Vectorization and NMF methods were the most successful approaches for extracting long-term trajectories. Using comprehensive multi-modal data helped us to develop an embedded space. This space was crucial for understanding the trajectories and dimensions in which the individuals traverse. Having this easily interpretable space, we were able to use a GMM unsupervised learning approach to identify new subtypes of the disorder based on disease progression. We also provided an in-depth analysis of these subtypes. Furthermore, we developed predictive models for early diagnosis, prognosis, and clinical trial stratification.

This work provides data-driven subtypes in distinct progression stages of PD and discusses an approach to predict the future rate of progression years from baseline using longitudinal clinical data. Predicting disease progression is a paramount challenge in treating and curing several complex diseases. This study is a step forward toward designing sophisticated machine-learning paradigms to facilitate the early diagnosis of PD progression and longitudinal biomarker discovery such as our finding of elevated Nfl in fast progressors (both at baseline and with regard to the rate of change per year). Predicting PD progression rates would lead to better patient-specific attention by recognizing the patients with a swift rate of progression at an early stage. The proposed disease progression and trajectory prediction algorithms can help healthcare providers to develop a methodical and organized course for clinical tests, which can be much more concise and effective in detection. These adaptations and modifications in clinics may help to diminish treatment and therapy costs for PD. Further, the capability to anticipate the trajectory of impending PD progression at the early stages of the disease is an advancement

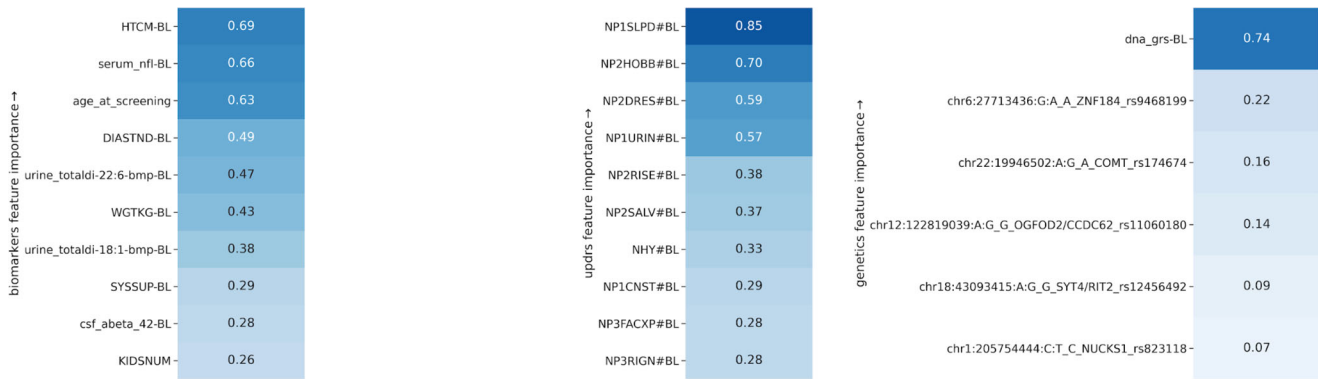


Fig. 8 Heatmap plot showing significant contributing clinical parameters (refer to Supplementary Table 6 for feature description) based on demographics, vital signs, baseline biospecimen, baseline MDS-UPDRS scores, and genetic measurements. The importance score of each feature is relative. BL baseline, HTCM height in cm, serum_nfl serum neurofilament light levels, age_at_screening Age at screening, DIASTND standing diastolic blood pressure (BP), urine_totaldi urine levels of di-22:6-bis (monoacylglycerol) phosphate, WGTKG weight in kg, SYSSUP supine systolic BP, csf_abeta_42 cerebrospinal fluid β -amyloid 1–42, KIDSNUM number of kids, dna_grs genetic risk score.

toward uncovering novel treatments for PD modification. The proposed analysis provides insights to inhibit or decelerate the progression of PD-related symptoms and subsequent deterioration in the characteristics of life that are accompanied by the disease.

METHODS

Study design and participants

This study included clinical data from the Parkinson's Progression Marker Initiative (PPMI, <http://www.ppmi-info.org/>) and the Parkinson's Disease Biomarkers Program (PDBP, <https://pdbp.ninds.nih.gov/>). Both cohort's data went through triage for missing data, 60-month assessment (36-month in PDBP), and comprehensive phenotype collection. Only data from participants with 60 months of follow-up for PPMI and 36 months for PDBP were included in the study. Overall, in the PPMI ($n = 294$ PD cases including 99 (34%) female; 154 controls including 58 (38%) female), and in the PDBP ($n = 263$ PD cases including 112 (43%) female; 115 controls including 64 (56%) female) passed the triage. The PPMI average age at the screening of PD cases was 61 ± 9.7 years and 60.3 ± 11 years for controls. The PDBP average age of PD cases was 64.3 ± 8.6 years and 63.6 ± 9.5 years for controls. The PPMI data also included 28 patients with other enrollments (10 PRODROMA; 8 GENPD; 6 GENU; 3 SWEDD; 1 REGPD), which were excluded. The PPMI and PDBP cohorts consist of observational data from comprehensively characterized PD patients and matched controls. All PD patients fulfilled the UK Brain Bank Criteria²¹. PD subjects enrolled in PPMI were drug naïve (i.e., not much treated with dopaminergic medications) for at least 2–3 years after enrollment. Being drug naïve is beneficial as we propose to build a disease progression tool for PD subtypes during early stages without complications from pharmacological interventions. Control subjects had no clinical signs suggestive of parkinsonism, no evidence of cognitive impairment, and no first-degree relative diagnosed with PD. Age and MDS-UPDRS Part III (objective motor symptom examination by a trained neurologist) distribution of cohorts at baseline were investigated using Kernel Density Estimation (KDE) to show that these independent cohorts are identically distributed and ensure the integrity of replication and validation (Supplementary Fig. 10, Supplementary Table 5). Each contributing study abided by the institutional review boards' ethics guidelines. All participants gave informed consent for inclusion in their initial cohorts and subsequent studies. Figure 1 provides an overview of the analyses and study design.

Dataset construction

The discovery and replication cohorts include visit data collected every 12 months starting from baseline to 60 months (36 months for PDBP) follow-up. In PPMI, visits at the 6 and 9-month time points from baseline were excluded in our analysis due to the high data missingness rate (>50%).

For each cohort, a comprehensive and shared set of longitudinally collected common clinical data elements were selected for analysis. Overall, 122 clinical features were available across six visits for PPMI (Supplementary Table 6) and 120 features across four visits for PDBP. We used the following features for the subtype identification stage:

- (i) International Parkinson's disease and Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part I, Part II, and Part III²²
- (ii) Cranial Nerve Examination (CN I-XII)
- (iii) Montreal Cognitive Assessment²³
- (iv) Hopkins Verbal Learning Test²⁴
- (v) Semantic Fluency test²⁵
- (vi) WAIS-III Letter-Number Sequencing Test²⁶
- (vii) Judgment of Line Orientation Test²⁷
- (viii) Symbol Digit Modalities Test²⁸
- (ix) SCOPA-AUT²⁹
- (x) State-Trait Anxiety Inventory for Adults³⁰
- (xi) Geriatric Depression Scale³¹
- (xii) Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease³²
- (xiii) REM-Sleep Behavior Disorder Screening Questionnaire³³
- (xiv) Epworth Sleepiness Scale³⁴.

In addition to these clinical measurements, biological and genetic-based features were included in the baseline subtype interpretation and the subtype prediction. These additional features include genotypes using imputed Illumina NeuroX array, vital signs, serum, CSF, and urine measurements. For genotyping data, we used the variants mapped to human genome build 38 (hg38) genotyping from unrelated European ancestry imputed genotype data passing standard QC metrics used to construct the genetic risk score (GRS)^{15–17}. The values indicate the number of copies of the minor allele of each variant for each subject. We used these values as categorical features. Patient characteristics include height, weight, blood pressure, and demographic details. For biological biomarkers, we assessed alpha-synuclein, total tau protein, β -amyloid 1–42 (A β 42), phospho-tau181 (p-Tau181) in CSF, serum neurofilament light (NfL), and urine levels of di-22:6-bis (monoacylglycerol) phosphate total in the urine. We studied the longitudinal variation of biomarkers and patients' characteristics

measurements across the identified PD subtypes. Furthermore, we investigated the biological measurements' role in discriminating PD subtypes.

Procedures and statistical analysis

The data analysis pipeline for this work was performed in Python (version 3.8) with the support of several open-source libraries (NumPy, pandas, matplotlib, seaborn, plotly, scikit-learn, UMAP, XGBoost, LightGBM, H2O, streamlit). To facilitate replication and expansion of our work, we have made the notebook publicly available on GitHub at <https://github.com/anant-dadu/PDProgressionSubtypes>. The code is part of the supplemental information; it includes the rendered Jupyter notebook with full step-by-step data preprocessing, statistical, and machine learning analysis. For readability, machine learning parameters have been described in the Python Jupyter notebook and not in the text of the paper. Our results are available on an interactive web browser (<https://anant-dadu-pdprogressionsubtypes-streamlit-app-aaah95.streamlitapp.com/>), which allows users to browse the PD progression space. In addition, the browser also includes predictive model interpretations allowing readers to explore feature contributions to model performance. For the streamlit website, we designed a surrogate XGBoost classification model for subtype prediction, which uses a single split with 70% training and 30% test data. The reported subtype classification performance in the manuscript is based on a more stringent nested loop procedure.

Data preprocessing

As the clinical features have varying directionalities, features were transformed to ensure the highest values uniformly represent the worst outcome while the lower value corresponds to greater health. To identify the features not following this pattern, we conducted a two-sample one-tailed t-test (null hypothesis: $\mu_{HC} > \mu_{PD}$) with cases versus controls as two samples. We transformed those features that showed a p -value of less than 0.05 (5% significance test). We further verified the transformations by manually reviewing the distribution of each feature. We performed feature clipping to minimize the influence of extreme outliers in the data. We limited all the features values between the range given by the second and 98th percentiles.

Only a few features had residual missingness that was distributed randomly across the patients at a rate of <5%. For these features, we performed data imputation using linear interpolation longitudinally (i.e., across visits) for each feature. Then, we transformed the dataset into a mathematically meaningful and naturally interpretable format. To achieve this objective, we (a) *vectorized* and (b) *normalized* all longitudinal data. Specifically, we first vectorized by transforming all observations of a particular parameter in a column vector, then appended all parameters together. We then used the *min-max* method to normalize the data.

Non-negative matrix factorization

Mathematically, NMF factorizes (deconstructs) the data into two matrices. Given a non-negative matrix $X \in R^{m \times n}$, a non-negative decomposition of the matrix X is a pair of non-negative matrices $U \in R^{m \times p}$ and $V \in R^{p \times n}$ such that $X = UV$. A large number of patient parameters are aggregated in a model that represents the underlying progression concept. In this particular use case of NMF, the matrix U contains the progression space latent vectors, and the second matrix V contains progression stand indicators corresponding to the latent vectors. Latent variables link observation data in the real world to symbolic data in the modeled world. By further looking into the matrix with progression space's latent vectors, we can identify the mapping and,

consequently, the implications (symbolic dimensions of the modeled progression space).

Latent space adjustment

The progression of space latent vectors (matrix U) learned by NMF shows some weight sharing among the projected dimensions. This weight sharing can be attributed to the presence of correlation in the data between symptomatology. To represent the progression space so that each progression indicator shows progression velocity for each symptom, we need to adjust the progression space. We performed transformation on NMF learned progression space by taking the weighted sum of the progression indicators. These weights represent the contribution of each dimension for distinct symptomatology.

Here, $C_{symptom\ type\ dimension\ (k)}^{symptom\ s}$ denotes the contribution of dimension k for features belonging to symptom s , $i_{dimension\ d}^{dimension\ d\ old}$ is the latent indicator learned by NMF, and $i_{dimension\ d}^{dimension\ d\ new}$ is the new indicator after the adjustment. The component contribution is calculated using the learned U matrix.

Through our use of NMF, we identified progressive features based on motor, cognitive, and sleep-based disturbances. Following this, unsupervised learning via Gaussian Mixture Models (GMM)³⁵ allowed the data to naturally self-organize into different groups relating to velocity of decline across these three categories, from non-PD controls representing normal aging to PD subtypes. GMM is a variant of mixture models, compared to other methods, the parametrization of a GMM allows it to efficiently capture products of variations in natural phenomena where the data is assumed generated from an independent and identically distributed (i.i.d.) mixture of Gaussian (normal) distributions. The assumption of normal distribution (and therefore, the use of GMM) is often used for population-based cohort phenomenon (Prentice 1986). We use the Bayesian Information Criterion (BIC) to select the number of PD clusters (subtypes)³⁶. The BIC method recovers the true number of components in the asymptotic regime (i.e., much data is available, and we assume that the data was generated i.i.d. from a mixture of Gaussian distributions). To replicate the subtype identification, we applied the GMM model developed in the PPMI data to an independent cohort with varying recruitment strategy and design: the PDBP cohort.

Unsupervised subtype identification

We used dimensionality reduction techniques to develop an interpretable representation of high modality longitudinal data. Dimensionality reduction techniques helped us to build the "progression space" where we can approximate each patient's position relative to both controls and other cases after the 60-month period in one-year intervals. We used the Non-negative Matrix Factorization (NMF) technique to achieve this aim^{37,38}. Alternative methods, such as principal component analysis and independent component analysis, did not perform as well as NMF on longitudinal clinical data due to the non-negative nature of our clinical tests. This process essentially collapses mathematically related parameters into the same multi-dimensional space, mapping similar data points close together.

Supervised early subtype prediction

After identifying progression classes using unsupervised learning, we built predictive models utilizing multiple supervised machine learning methods, including the *ensemble learning* approach. This method combines multiple learning algorithms to generate a better predictive model than could be obtained using a single learning algorithm³⁹. To do this, we used stacking ensembles of three supervised machine learning algorithms (Random forest⁴⁰,

LightGBM⁴¹, and XGBoost⁴²) to predict PD clinical subtypes using the data obtained at the time a neurologist first reviewed the patient as the input (combining baseline and varied time points). This approach outperformed other methods in preliminary testing, such as support vector machines (SVM) and simple lasso-regression models. Besides the predictive performance, we chose an ensemble approach due to the nature of our data and problem: (i) decision trees are intrinsically suited for multiclass problems, while SVM is intrinsically two-class, (ii) they work well with a mixture of numerical, categorical, and various scale features, (iii) they can be used to rank the importance of variables in a classification problem and in a natural way which helps the interpretation of clinical results, and (iv) it also gives us the probability of belonging to a class, which is very helpful when dealing with individual subject progression prediction. We developed three predictive models to predict the patient's progression class after 60 months based on varying input factors: (a) from baseline clinical factors, (b) from baseline and first-year clinical factors, and (c) from biomarkers and genetic measurements.

To validate the effectiveness of our predictive models, we used a nested cross-validation (CV) approach with 5 folds in both inner and outer loops. Specifically, we randomly divided the dataset into five subsamples (outer folds). Each of the subsamples was used as the testing data exactly once, while the remaining (training) data was used for hyperparameter tuning and model training. The hyperparameters were chosen based on their average performance on training data during the inner cross-validation loop. The workflow of the approach is depicted in Supplementary Fig. 11. The performance of the algorithm was measured by the area under the receiver operating curve (AUC) generated by plotting sensitivity vs. (1 – specificity). We used a macro-average AUC score computed by averaging the metric independently for each class (hence treating all classes equally for predicting fast, moderate, and slow progressing cases). The five results from the multiple iterations were averaged to produce a single estimation of performance across these three classes.

To conclusively validate the algorithm, we also evaluated the performance of the predictive models (trained on the PPMI measurements) on the independent PDBP cohort. To replicate, we trained the supervised model on PPMI latent weights (at baseline) and then used the same model on the PDBP latent weights (at baseline). We show that the predictive models preserve their high accuracy applied to another dataset.

Biomarker based prediction

There were 448 observations in total. We did not include plasma, CSF hemoglobin, and CSF glucosylceramide features because of their high missing data (>35%). Baseline CSF data were missing for p-tau in 51 patients, for total-tau in 26 patients, for abeta-42 in 20 patients, and alpha syn in 15 patients. Serum NfL is missing in 22 participants, 17 participants did not have data for DNA GRS scores, and 36 participants had missing APOE status. The overall missing percentage for the above measurements was ~5%. We imputed the missing predictor variable data with means for numerical features and used most frequent category for categorical features. The remaining features include demographic information (education year, biological sex, birthdate, race), vital signs (weight, height, blood pressure), and family history (parents' PD status) with no patients having missing data for them. For hg-38 genetic measurements, data was missing for 8 participants. We removed these patients from our study. Finally, we had 440 participants and 39 and 64 features for biomarkers and genetics measurements, respectively. All the genetic features are considered as categorical except DNA GRS. For the combined model, we concatenated both the biospecimen and genetic features. This concatenated vector was used as input for the classification of PD subtypes.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data used in this study was access-controlled from the Parkinson's Progression Marker Initiative (PPMI, <http://www.ppmi-info.org/>) and the Parkinson's Disease Biomarkers Program (PDBP, <https://pdbp.ninds.nih.gov/>), and require individual sign-up to access the data. Additionally, we have developed an interactive website [<https://anant-dadu-pdprogressionsubtypes-streamlit-app-aaah95.streamlitapp.com/>] where researchers can investigate components of the predictive model and can investigate feature effects on a sample and cohort level.

CODE AVAILABILITY

To facilitate replication and expansion of our work, we have made the notebook publicly available on GitHub at [<https://github.com/anant-dadu/PDProgressionSubtypes>]. It includes all code, figures, models, and supplements for this study. The code is part of the supplemental information; it includes the rendered Jupyter notebook with full step-by-step data preprocessing, statistical, and machine learning analysis.

Received: 4 August 2022; Accepted: 29 November 2022;
Published online: 16 December 2022

REFERENCES

- Hughes, A. J., Daniel, S. E., Kilford, L. & Lees, A. J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry* **55**, 181–184 (1992).
- Postuma, R. B. et al. MDS clinical diagnostic criteria for Parkinson's disease. *Mov. Disord.* **30**, 1591–1601 (2015).
- Stebbins, G. T. et al. 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. *Mov. Disord.* **28**, 668–670 (2013).
- Jankovic, J. et al. Variable expression of Parkinson's disease: a base-line analysis of the DAT ATOP cohort. *Neurology* **40**, 1529–1529 (1990).
- Zetuskay, W. J., Jankovic, J. & Pirozzolo, F. J. The heterogeneity of Parkinson's disease: clinical and prognostic implications. *Neurology* **35**, 522–526 (1985).
- Faghri, F. et al. Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study. *Lancet Digit Health* **4**, e359–e369 (2022).
- Satone, V. K. et al. Predicting Alzheimer's disease progression trajectory and clinical subtypes using machine learning. Preprint at <https://www.biorxiv.org/content/10.1101/792432v2> (2020).
- van Rooden, S. M. et al. The identification of Parkinson's disease subtypes using cluster analysis: a systematic review. *Mov. Disord.* **25**, 969–978 (2010).
- Fereshtehnejad, S.-M. et al. New clinical subtypes of Parkinson disease and their longitudinal progression: a prospective Cohort comparison with other phenotypes. *JAMA Neurol.* **72**, 863–873 (2015).
- 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).
- Lawton, M. et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. *J. Neurol. Neurosurg. Psychiatry* **89**, 1279–1287 (2018).
- Dadu, A. *Application of Machine Learning to the Detection and Prediction of Parkinson's Disease Subtypes*, University of Illinois Urbana-Champaign. <http://hdl.handle.net/2142/113054> (2021).
- Faghri, F. et al. Predicting onset, progression, and clinical subtypes of Parkinson disease using machine learning. Preprint at <https://www.biorxiv.org/content/10.1101/338913v2> (2018).
- Nalls, M. A. et al. Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. *Lancet Neurol.* **14**, 1002–1009 (2015).
- 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. *Lancet Neurol.* **18**, 1091–1102 (2019).
- Makarious, M. B. et al. GenoML: automated machine learning for genomics. *arXiv* <https://doi.org/10.48550/arXiv.2103.03221> (2021).

17. Makarios, M. B. et al. Multi-modality machine learning predicting Parkinson's disease. *npj Parkinson's Dis.* **8**, 1–13 (2022).
18. Zhang, X. et al. Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study. *Sci. Rep.* **9**, 797 (2019).
19. Krishnagopal, S., Coelln, Rvon, Shulman, L. M. & Girvan, M. Identifying and predicting Parkinson's disease subtypes through trajectory clustering via bipartite networks. *PLoS One* **15**, e0233296 (2020).
20. Marras, C. & Chaudhuri, K. R. Nonmotor features of Parkinson's disease subtypes. *Mov. Disord.* **31**, 1095–1102 (2016).
21. Sudlow, C. et al. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
22. 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. *Mov. Disord.* **23**, 2129–2170 (2008).
23. Nasreddine, Z. S. et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).
24. Brandt, J. The hopkins verbal learning test: development of a new memory test with six equivalent forms. *Clin. Neuropsychol.* **5**, 125–142 (1991).
25. Goodglass, H., Kaplan, E. & Barresi, B. *The Assessment of Aphasia and Related Disorders*. 3rd edn (Lippincott Williams & Wilkins, 2001).
26. Wechsler, D. *WAIS-III: Wechsler Adult Intelligence Scale* 3rd edn (Psychological Corporation, Seattle, 1997).
27. Benton, A. L., Varney, N. R. & Hamsher, K. D. Visuospatial judgment. A clinical test. *Arch. Neurol.* **35**, 364–367 (1978).
28. SMITH & A. in *Learning Disorders* Vol. 341, 83–91 (Special Child Publications, 1968).
29. Visser, M., Marinus, J., Stiggelbout, A. M. & Van Hilten, J. J. Assessment of autonomic dysfunction in Parkinson's disease: The SCOPA-AUT. *Mov. Disord.* **19**, 1306–1312 (2004).
30. Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R. & Jacobs, G. A. State-trait anxiety inventory for adults. *APA PsycTests* <https://doi.org/10.1037/06496-000> (1983).
31. Yesavage, J. A. & Sheikh, J. I. 9/Geriatric depression scale (GDS) recent evidence and development of a shorter version. *Clin. Gerontol.* **5**, 165–173 (1986).
32. Weintraub, D. et al. Validation of the questionnaire for impulsive-compulsive disorders in Parkinson's disease. *Mov. Disord.* **24**, 1461–1467 (2009).
33. Stiasny-Kolster, K. et al. The REM sleep behavior disorder screening questionnaire—a new diagnostic instrument. *Mov. Disord.* **22**, 2386–2393 (2007).
34. Johns, M. W. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep* **14**, 540–545 (1991).
35. McLachlan, G. J. & Basford, K. E. *Mixture Models: Inference and Applications to Clustering* Vol. 84 (Marcel Dekker, 1988).
36. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
37. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
38. Lee, D. D. & Seung, H. S. in *Advances in Neural Information Processing Systems* Vol. 13 (eds. Leen, T. K., Dietterich, T. G. & Tresp, V.) 556–562 (MIT Press, 2001).
39. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **33**, 1–39 (2010).
40. Breiman, L. Random forests. *Mach. Learn* **45**, 5–32 (2001).
41. Ke, G. et al. in *Advances in Neural Information Processing Systems* Vol. 30 (eds. Guyon, I. et al.) 3146–3154 (MIT Press, 2017).
42. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proc. 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).

ACKNOWLEDGEMENTS

We thank the patients and their families who contributed to this research. This research was supported in part by the Intramural Research Program of the National Institute on Aging (NIA) and National Institute of Neurological Disorders and Stroke (NINDS), both part of the National Institutes of Health, within the Department of Health and Human Services; project number ZIA AG000534, ZO1 AG000949, ZIANS003154, and the Michael J Fox Foundation. This work has also been supported in part through the grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH big data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). Data used in the preparation of this article were obtained from the Parkinson's 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 Parkinson's Research and funding

partners, including Abbvie, Avid Radiopharmaceuticals, Biogen Idec, Bristol-Myers Squibb, Covance, Eli Lilly & Co., F. Hoffman-La Roche, Ltd., GE Healthcare, Genentech, GlaxoSmithKline, Lundbeck, Merck, MesoScale, Piramal, Pfizer, and UCB. Data and biospecimens used in the preparation of this manuscript were obtained from the Parkinson's Disease Biomarkers Program (PDBP) Consortium, part of the National Institute of Neurological Disorders and Stroke at the National Institutes of Health. Investigators include: Roger Albin, Roy Alcalay, Alberto Ascherio, DuBois Bowman, Alice Chen-Plotkin, Ted Dawson, Richard Dewey, Dwight German, Xuemei Huang, Rachel Saunders-Pullman, Liana Rosenthal, Clemens Scherzer, David Vaillancourt, Vladislav Petyuk, Andy West and Jing Zhang. The PDBP Investigators have not participated in reviewing the data analysis or content of the manuscript.

AUTHOR CONTRIBUTIONS

A.D., A.B.S., M.A.N., R.H.C., and F.F. contributed to the concept and design of the study. A.D., V.S., R.K., H.L., H.I., M.B.M., K.J.B., S.B.-C., L.J.S., A.J.N., C.B., M.A.N., and F.F. were involved in the acquisition of data, data generation, and data cleaning. A.D., V.S., R.K., S.H.H., M.A.N., and F.F. did the analysis and interpretation of data. A.D., V.S., R.K., S.H.H., H.L., H.I., M.B.M., K.J.B., S.B.-C., L.J.S., A.J.N., A.I. D., C.B., K.M., S.W.S., A.B.S., M.A.N., R.H.C., and F.F. contributed to the drafting of the article and revising it critically.

FUNDING

Open Access funding provided by the National Institutes of Health (NIH).

COMPETING INTERESTS

A.D., H.L., H.I., M.A.N., and F.F.'s declare no competing non-financial interests but the following competing financial interests as their participation in this project was part of a competitive contract awarded to Data Tecnica International LLC by the National Institutes of Health to support open science research. M.A.N. also currently serves on the scientific advisory board for Character Bio and is an advisor to Neuron23 Inc. The study's funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report. Authors V.S., R.K., S.H.H., M.B.M., K.J.B., S.B.-C., L.J.S., A.J.N., Ali D., C.B., K.M., S.W.S., A.B.S., and R.H.C. declare no competing financial or non-financial interests. All authors and the public can access all data and statistical programming code used in this project for the analyses and results generation. F.F. takes final responsibility for the decision to submit the paper for publication.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41531-022-00439-z>.

Correspondence and requests for materials should be addressed to Faraz Faghri.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2022