RESEARCH ARTICLE

Identifying and predicting Parkinson's disease subtypes through trajectory clustering via bipartite networks

Sanjukta Krishnagopal[1]*, **Rainer von Coelln**[2], **Lisa M. Shulman**[2], **Michelle Girvan**[1,3]

1 University of Maryland College Park, College Park, MD, United States of America, **2** University of Maryland School of Medicine, Baltimore, MD, United States of America, **3** Santa Fe Institute, Santa Fe, New Mexico, United States of America

* sanju33@gmail.com

Abstract

Chronic medical conditions show substantial heterogeneity in their clinical features and progression. We develop the novel data-driven, network-based Trajectory Profile Clustering (TPC) algorithm for 1) identification of disease subtypes and 2) early prediction of subtype/ disease progression patterns. TPC is an easily generalizable method that identifies subtypes by clustering patients with similar disease trajectory profiles, based not only on Parkinson's Disease (PD) variable severity, but also on their complex patterns of evolution. TPC is derived from bipartite networks that connect patients to disease variables. Applying our TPC algorithm to a PD clinical dataset, we identify 3 distinct subtypes/patient clusters, each with a characteristic progression profile. We show that TPC predicts the patient's disease subtype 4 years in advance with 72% accuracy for a longitudinal test cohort. Furthermore, we demonstrate that other types of data such as genetic data can be integrated seamlessly in the TPC algorithm. In summary, using PD as an example, we present an effective method for subtype identification in multidimensional longitudinal datasets, and early prediction of subtypes in individual patients.

Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder, affecting an estimated 7-10 million people worldwide [1]. The cause of PD is unknown, and the disease course is variable with age of onset and rate of progression differing across the population [2]. Furthermore, the clinical presentation is variable, with a broad range of possible motor and non-motor symptoms [3]. Based on these differences, multiple PD subtypes have been proposed, based on clinical intuition or unbiased data-driven approaches like cluster analysis [4]. Disease subtypes, which are likely to differ by the underlying etiology, treatment responsiveness and prognosis, will therefore facilitate PD research, management, and counseling of patients regarding prognosis [5, 6].

There is currently no consensus on Parkinson's subtypes that are biologically valid and clinically relevant, and the best approach for identifying such subtypes remains elusive [7]. Lack of

integration of longitudinal data for a large number of variables and lack of data-based prognoses are limitations of existing approaches [8].

Network medicine [9–13] offers a promising approach for untangling the complexities due to multiple influences on disease manifestation and progression via analysis of interconnections within data. For example, studies of the human disease network (i.e. the 'diseaseome') [13], in which diseases are linked if they share one or more associated genes, are useful for identifying disease pathways and predicting other disease-related genetic variants [11]. With few exceptions, most network medicine studies have focused on biomolecular data [13–16] rather than the complexities of clinical phenotypic assessments, and disease subtyping based on disease progression patterns is relatively unexplored [17, 18]. Another possible benefit of network medicine approaches is that they offer ways to integrate different types of data, for example to simultaneously incorporate clinical assessments with genetic data. This is especially important for PD, as a large number of genetic variants have been identified as risk factors [19]. Further, evidence has emerged that the same genetic risk variants also determine certain clinical features of the disease, highlighting the need to explore novel approaches that integrate genetic data into clustering (or subtyping) algorithms [20, 21].

Technological innovations in data processing and storage capacity have enabled development of large clinical datasets, containing longitudinal clinical and biological data. In this work we use data from the Michael J. Fox Foundation's Parkinson's Progression Markers Initiative (PPMI), a worldwide study to establish a comprehensive set of clinical, imaging and genetic data (http://www.ppmi-info.org). Such datasets require sophisticated data-driven approaches for effective extraction and analysis of clinically relevant information. Data-driven methods are typically applied to diseases in two ways: disease-specific, i.e., identifying disease subtypes and variable progression patterns from large scale patient data, and patient-specific, i.e., predicting disease subtype and trajectory in the individual patient based on their data. Our work incorporates both these perspectives and presents a network science method that not only identifies disease subtypes using diverse types of patient data (e.g., genetic and clinical variables), but is also predictive. We present our results based on a PD dataset, however this method is easily applied to other chronic medical conditions.

To provide an intuitive data-driven solution that is both disease- and patient-centric, we develop the novel Trajectory Profile Clustering (TPC) algorithm to identify PD subtypes through similarities in patterns of progression. Additionally, we demonstrate the predictive ability of our algorithm on a test/validation cohort of new patients. We also explore inclusion of four PD genetic variants in our approach, to demonstrate its capacity to simultaneously incorporate clinical, demographic, and genetic information. Thus, TPC is a data-driven algorithm that can incorporate different types of data (e.g., genetic, clinical etc.) and different weighting schemes for different variables in order to cluster patients according to the similarity of their disease progression. In addition, TPC also offers predictive power, making it a useful tool for clinicians in the study of multivariate, progressive disease datasets. Our method, to the best of our knowledge, presents a new and easily generalizable approach for robust subtype identification by accounting for disease progression patterns in addition to overall variable profiles. This work is aimed at bridging the gap between the computational methodologies developed by network and data scientists and the clinical experience of health professionals.

Materials and methods

Description of data

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). The data consists of patient

Domain	Scale/Variable
Demographics	
Gender	Gender
Age	Age
General PD Severity	
MDS-UPDRS1	Movement Disorders Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) - Part 1
MDS-UPDRS2	Movement Disorders Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) - Part 2
MDS-UPDRS3	Movement Disorders Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) - Part 3
T-MDS-UPDRS	Total Movement Disorders Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS)
Cognitive	
JOLO	Benton Judgement of Line Orientation
SDM	Symbol Digit Modalities Test
MoCA	Montreal Cognitive Assessment
HVLT	Hopkins Verbal Learning Test
LNS	Letter Number Sequencing
SFT	Semantic Fluency Test
Disability	
SEADL	Schwab and England Activities of Daily Living
Sleep	
RBDQ	Rapid Eye Movement Sleep Behavior Disorder Questionnaire
ESS	Epworth Sleepiness Scale
Autonomic	
SCOPA-AUT	Scales for Outcomes in Parkinson's Disease - Autonomic
Mental Health	
GDS	Geriatric Depression Scale
STAI	State -Trait Anxiety Inventory
Genetic Risk Loci	
G1	rs11060180
G2	rs6430538
G3	rs823118
G4	rs356181

Fig 1. Description of PPMI Data. Data includes two demographic variables, outcome variables from six clinical domains, and four genetic single nucleotide polymorphisms.

https://doi.org/10.1371/journal.pone.0233296.g001

variable values across 5 time points: baseline values (which we denote as year 0) and years 1,2,3, and 4. Of the 430 patients at baseline in this dataset, 314 patients remained in year 4. Once patients with incomplete data were excluded, 194 patients remained in our analysis. Twenty percent of this population (number of individuals n = 39) was kept as a test/validation dataset. The remainder of the patients (n = 155) formed the training dataset that was used in the algorithm to identify PD subtypes. The data included demographics (gender and age in year 4), clinical variables from six clinical domains (General PD Severity, Disability, Cognition, Autonomic Function, Sleep, and Mental Health) and 4 PD genetic variants (Fig 1). PPMI motor assessment was performed in a 'practically defined off' state, i.e., subjects are asked to withhold their medication prior to the assessment for 12 hours for a defined "OFF" medication score, practically eliminating medication effects on motor symptoms in this dataset.

Trajectory Profile Clustering algorithm

Our Trajectory Profile Clustering algorithm is designed to group together patients based on the similarities of their disease trajectories. The algorithm proceeds as follows:

1. *Create bipartite networks connecting individuals to variables*: At time point t (e.g., baseline, year 1, year 2, etc.) we construct an $N \times V$ bipartite graph modeling connections between individuals and disease variables, where N is the number of individuals in the training population and V is the total number of variables, as illustrated in Fig 2. For M time points, we can represent the set of these bipartite graphs as an $N \times V \times M$ multidimensional array, where X_{ivt} gives the value of individual i's disease variable v at time t.

2. *Transform data for variable uniformity*: For each non-binary variable, we determine its 'direction'. For variable v, if higher values of the variable are associated with greater disease severity then its direction $d_v = +1$; otherwise $d_v = -1$. For our data, clinical variables ESS,
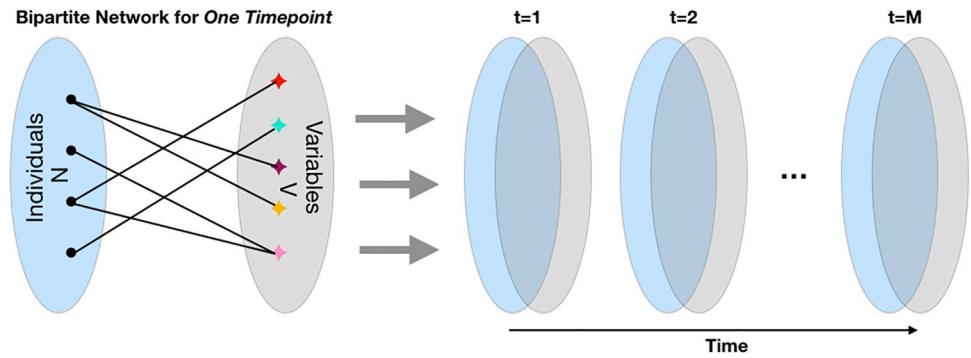
Fig 2. Stacking bipartite networks across time. An illustration of an individual-variable bipartite graph at one timestep (left). Set of bipartite graphs across time (right).

RBDQ, GDS, STAI, UPDRS and age have $d_v = 1$, and HVLT, JOLO, SFT, LNS, SDM, MoCA, SEADL have $d_v = -1$. We then define a new $N \times V \times M$ multi-dimensional array Y such that $Y_{ivt} = d_v X_{ivt}$ for non-binary variables. For binary variables, $Y_{ivt} = X_{ivt}$.

3. *Construct patient trajectory profiles*: For each patient i, we construct a $V \times M$ trajectory profile matrix, T^i. The matrix entries of T^i are calculated as follows:

- For non-binary variables:

$$T^i_{vt} = 1 \text{ if } Y_{ivt} > \theta_v$$
$$= 0 \text{ otherwise.}$$

(1)

where θ_v is the threshold for variable v. In this manuscript, we set θ_v to the median baseline value of variable v in the training data. We threshold the connections, i.e., the individual is only connected with disease variables for which they have a high enough severity. This thresholding causes patients to be shown as unconnected to all variables in Fig 2.

- For binary variables:
  For gender: $T^i_{vt} = 1$ if the patient is male, $T^i_{vt} = 0$ otherwise. *For genetic risk loci*: $T^i_{vt} = 1$ if patient contains single nucleotide polymorphisms (SNP) v, $T^i_{vt} = 0$ otherwise. Each SNP is treated as independent.

4. *Create a patient-patient network with connections based on trajectory similarity*: After having defined the trajectory profile matrix T^i for each individual i, we create a patient-patient network P of all patients in the training set. The nodes of this network correspond to patients and the strength of a link between patient i and patient j captures the similarity of their trajectory profiles. P has an adjacency matrix given by:

$$P_{ij} = \sum_{v,t} (T^i_{vt} \equiv T^j_{vt}).$$

(2)

In other words, P_{ij} gives the number of matrix entries for which trajectory profile T^i has the same value as T^j. This formulation implies that variables are equally weighted. Other applications may require unequal weighting for variables and time points, in which case one may calculate the patient-patient matrix as follows: $P_{ij} = \sum_{v,t} w_{vt} (T^i_{vt} \equiv T^j_{vt})$ where w_{vt} is the weight of variable v at time t. An alternate more finely resolved approach to constructing the patient-patient network P would be, for example, to divide the baseline data for each variable into quartiles. In this case, the strength of a link between two patients would be

determined by the number of variable-timesteps for which their values landed in the same quartile. In preliminary investigations, the more finely resolved approach gave similar results, so for simplicity we focus in this manuscript on the median-based discretization scheme defined by Eq 1.

5. *Cluster the network to identify communities/subtypes*: We then perform Louvain community detection [22] to maximize the Newman-Girvan modularity function [23] on the uni-partite network defined by the weighted matrix P. As is common in network community detection approaches [24], the number of communities is not set a priori, but rather chosen so that the modularity is maximized. This process allows us to cluster trajectory profiles, and hence patients, into communities (subtypes) which are relatively densely connected.

6. *Construct aggregate profiles to characterize each community/subtype*: We average the trajectory profiles of all patients in each community C^l to obtain the 'community/subtype profile' S^l. The subtype profile is indicative of the variable features that describe the subtype. More specifically, it is the normalized average of the trajectory profiles of all the patients in that subtype, i.e., S^l is a $V \times M$ matrix with elements defined by

$$S_{vt}^l = \frac{\sum_{i \in C^l} T_{vt}^i}{N_l U_{v0}}$$ (3)

where N_l is the total number of individuals in community C^l. U_{v0} is a normalization constant that represents the average value for variable v in the baseline: $U_{v0} = \frac{\sum_i T_{v0}^i}{N}$, and 0 denotes the baseline year.

Prediction scheme for test patients

From baseline data, we predict the community/subtype that an individual test patient (patient whose data was not used in identifying the PD subtypes) belongs to. We then check whether the test patient is still aligned with the same community/subtype after 4 years to demonstrate the utility of our baseline prediction.

To predict test patient i's subtype from his/her baseline profile, we find the community (subtype) C^l whose baseline community profile, with elements S_{v0}^l, has the smallest Euclidean distance from the patient's baseline profile. In other words, l is chosen to minimize the distance

$$d_0^{il} = \sqrt{\sum_v (T_{v0}^i - S_{v0}^l)^2}.$$ (4)

Does the patient's trajectory match the subtype's trajectory? We then investigate the quality of the subtype/community baseline prediction at a later time t by calculating the patient's subtype/community $C^{l'}$ is chosen to minimize the distance between the community profile and the patient's profile at time t:

$$d_t^{il'} = \sqrt{\sum_v (T_{vt}^i - S_{vt}^{l'})^2}.$$ (5)

The prediction accuracy is then defined as the fraction of test patients for which the subtype identification (l) from the baseline matches the subtype identification (l') at a later time t.

Results

TPC algorithm for PD subtype identification

In this section, we present the disease subtypes (communities) identified by our method from the training patient data. Maximizing Newman-Girvan modularity on the patient-patient trajectory profile network gives us three distinct subtypes, i.e., three is the optimal number of subtypes for this data, as indicated by the modularity measure.

The clinical profiles plus demographics of each subtype as compared to the entire study population are shown for baseline and years 1-4 in Fig 3. The darkness of the shade of grey of a continuous variable in a year denotes the fraction of the subtype population that has a value above the median of the total population baseline for that variable. The darkness of the shade of grey for a binary variable is the fraction of the subtype population containing that variable (male in the case of the variable gender). In the raw data, a higher raw score in some variables (such as the Montreal Cognitive Assessment) implies a healthier/less severely affected patient, while for other scales, the opposite is true (higher score = greater severity). Therefore, in step one of our algorithm we transformed the data, so that for all variables except for the genetic and demographic variables, a higher score is associated with greater severity of that variable and a deeper shade of grey.
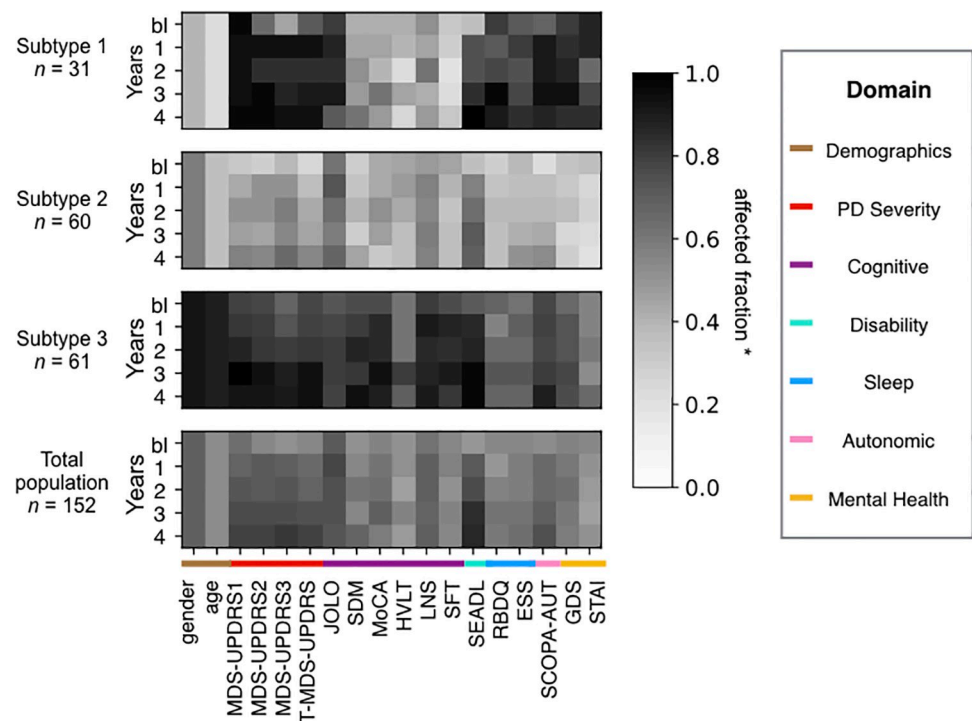
Fig 3. Variable profiles of the Parkinson's subtypes identified by the TPC algorithm. Subtypes/communities identified by our algorithm: top three panels show three subtype/community profiles (average profile of all patients in the subtype). Subtypes identified by the algorithm containing fewer than 10 patients are not shown (3 patients fall under this category). The bottom panel shows the total population profile. The shade of grey indicates the affected fraction, i.e, fraction above baseline median in the direction of disease progression for the continuous variables, and fraction that is male for gender. n is the number of patients in the subtype. The variable names are listed below the panels (See Fig 1 for description).

https://doi.org/10.1371/journal.pone.0233296.g003

Description of the subtypes

As seen in Fig 3, the subtypes can be outlined as follows:

- Subtype 1 is the 'mixed subtype', with a striking discrepancy between severe impairment of motor and autonomic function, mental health and sleep on the one hand, and good cognition on the other hand, both at baseline and over time, and also young and predominantly female;

- Subtype 2 is the 'mild subtype', with milder than average impairment in all domains (motor, cognitive, autonomic and mental) at baseline and throughout the study duration (age and gender distribution close to the average of the entire population);

- Subtype 3 is the 'severe subtype', with worse than average impairment in all domains, in particular motor and cognition. This subtype is also predominantly male and older than average. Autonomic and mental function is less impaired than in the mixed subtype (subtype 1)

The bottom panel in Fig 3 shows the profile of the total population. Since the threshold variable severity in an individual is set with respect to the median of the total population at baseline, the total population baseline profile for all variables has a value close to 0.5 (i.e., 50% of the total population at baseline has a value of 1 for any variable, and the other half has a value of 0). Fluctuations of the baseline total population value around 0.5 occur when multiple people in the population have a value coinciding with the baseline median. In the Appendix A, we provide statistical analyses comparing the subtypes at the baseline and the final timepoint (year 4). These analyses support the subtype descriptions provided above.

Early prediction of patient subtypes

In addition to identifying PD subtypes, our method predicts the individual patient subtype years in advance. In this section we use the test patient cohort (n = 39) to assess the accuracy of early prediction of disease subtype. Data from these test patients was not used in the identification of the subtypes. Fig 4 shows the prediction of future PD subtype based on baseline data for 39 test patients that run across the horizontal axis. The top panel shows the Euclidian distance between the baseline profile of a patient and the baseline profile of each subtype (subtypes are shape coded). The subtype with which the patient has minimum *baseline* distance is the 'predicted subtype', and is marked in red. Patients are organized from left to right in order of decreasing confidence, i.e., from minimum to maximum distance of the patients' baseline profile with the predicted subtype baseline profile. The remaining panels follow the same plotting scheme for consecutive years. The red-coding of baseline predicted community makes it easy to track across the years. Finally, in year 4, we assess the accuracy of our predictions by identifying the 'actual final subtype' (the subtype with minimum distance to the patient in year 4). If the actual final subtype and the predicted subtype are the same, then we consider our prediction to be successful for that patient. In other words, for a patient, if in year 4 the red subtype has the minimum distance (is below the black subtypes) then our prediction is successful. For the test PD patients in the PPMI dataset, our algorithm uses only their baseline year data to predicts their PD subtype after 4 years of disease progression with 72% accuracy.

Incorporating genetic data into the TPC algorithm

Genetic variants are increasingly recognized as important determinants of disease subtype and disease progression and prognosis. As an exploratory objective, we investigated the integration of genetic variants (single nucleotide polymorphisms, or SNPs) in previously identified PD risk loci into our TPC-based approach. Each patient has 2 copies for each piece of genetic
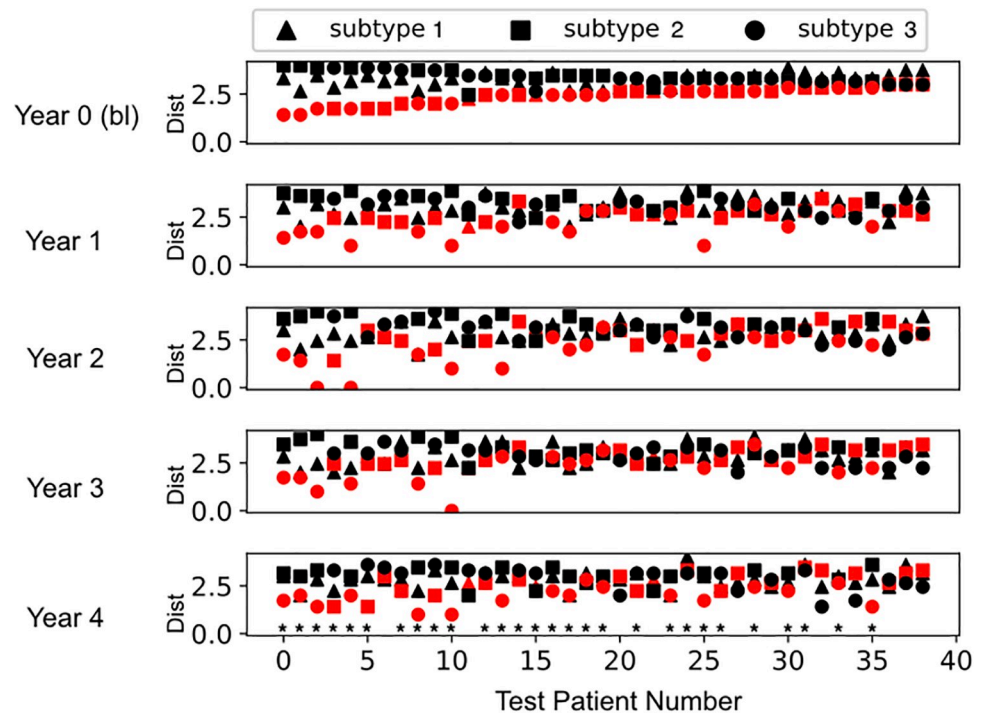
Fig 4. Prediction of test patients into the subtypes. The ith panel (row) shows the distance between the test patient i^{th} year profile and the ith year subtype profile (shape coded). The predicted subtype for each individual (subtype with minimum baseline-year distance) is colored red to allow for tracking across the years (panels). Prediction accuracy in year 4 is 72%. Patients whose year 4 subtype is correctly predicted from their baseline data are designed by a star. Data includes 39 test patients and 18 clinical variables across 5 time points: baseline (bl) or year 0 + years 1,2,3,4.

https://doi.org/10.1371/journal.pone.0233296.g004

information, and there are by definition 2 variants for each SNP. Hence, there are 3 possible combinations of the 2 variants for each of the genetic risk loci. PPMI contains information for 28 such SNPs for each patient. As a proof of principle, we selected 4 of those 28 SNPs to be included in our approach. For one of those SNPs (rs356181/2, labeled 'G4' in our study), an association with PD motor has recently been described [25], making this an obvious choice for our study. Recently, genotype-phenotype correlations have been described for a number of SNPs associated with PD risk [21, 26, 27]. However, there was minimal overlap in terms of which genetic variants were associated with specific clinical features of PD, even though two of these studies were performed by the same consortium, analyzing data from essentially the same collection of large PD cohorts [21, 27]. Consistent with our proof-of-principle approach, we therefore picked the additional 3 SNPs based on their high minor allele frequency, so that all 3 possible combinations of the 2 genetic variants were present in sufficient numbers in our study population of 194 subjects. The G allele of one of these SNPs (rs1106180, labeled 'G1' in our study) is associated with a later age of onset [21].

Fig 5 shows the five subtypes identified when genetic data is introduced. Here, the number of subtypes is determined by maximizing the network modularity measure for the network created using clinical and genetic information. The plots Fig 5(a) and 5(b) are organized in the same way as Figs 3 and 4 respectively. In Fig 5(a), the darkness of the shade of grey of a variable in a year denotes the fraction of the subtype that has a value above the baseline median of the total population of that variable, and the color coding on the x axis denotes the domains as in Fig 3.
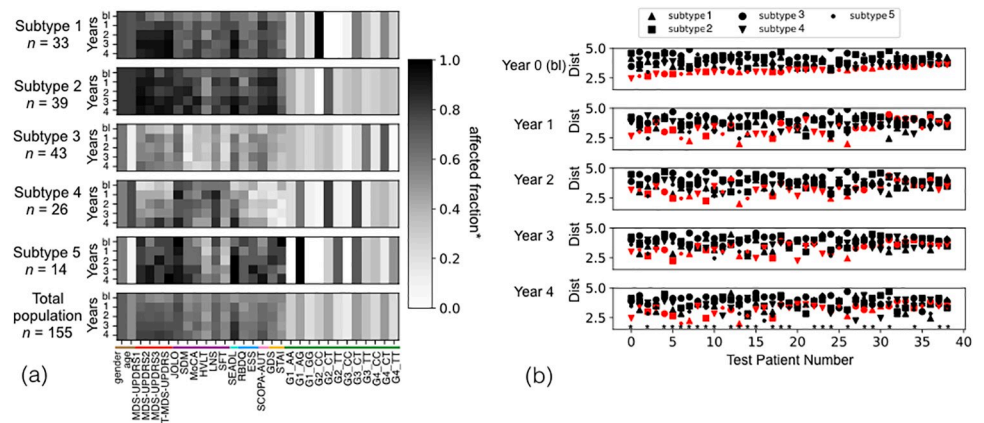
Fig 5. Variable profiles and test patient subtype prediction using clinical and genetic data. (a) Top five panels show five average community (subtype) profiles, identified by our TPC algorithm. The bottom panel shows the total population profile. The legend is a measure of the affected fraction, i.e, fraction above baseline median in the direction of disease progression for the continuous variables, and fraction that is male and fraction containing the genetic SNP for gender and genetic variables respectively. n is the number of patients in the community. (b) The mth panel shows the distance between the test patient m^{th} year profile and the mth year profile of the subtypes (shape coded). The predicted subtype for each individual (subtype with minimum baseline distance) is colored red to allow for tracking across the years (panels). Prediction accuracy in year 4 is 67%. Patients whose year 4 subtype is correctly predicted from their baseline data are designed by a star. Data includes 39 test patients and 18 clinical variables across 5 time points: baseline (bl) or year 0 + years 1,2,3,4.

Subtypes 1 (top) and 2 have relatively similar clinical profiles, with the difference being in their genetic profile. The CC genotype of G2, and CT genotype of G3 and G4 are frequent in subtype 1, and the CT is the most common genotype of G2 in subtype 2. Subtypes 3 and 4 relatively mild symptoms at baseline, and an overall benign progression of most variables over the course of 4 years. Subtype 3 includes more patients who are younger and less cognitively impaired than average. Subtype 4 has less psychiatric, autonomic and sleep impairments but intermediate motor impairments (MDS-UPDRS3). Finally, subtype 5 is small ($n = 14$) and young with rather severe symptoms at baseline and rapid progression across most clinical domains. Each of the subtypes has a distinct genetic profile. When genetic data is added to the analysis, baseline prediction of patient subtype in the test group 4 years later shows an accuracy of 67%.

Discussion

Multidimensional clinical datasets are valuable resources that are not used to their full potential due to the analytic challenges of diverse biomarkers and outcome variables. We describe development of a method to identify disease subtypes based on the pattern of progression of multidimensional clinical data including demographics, clinical variables, and genetics. We then validate our method by measuring the accuracy of subtype prediction in individual patients based on baseline clinical and genetic variables. The disease subtypes are characterized by patterns of progression of the clinical variables. The concordance between our results with the domain-structure of the variables supports our approach. For example, in the clinical-only case, subtypes 1 and 3 have high progression of all PD severity variables and subtype 2 has a low progression of all PD severity variables. Variables within other domains such as Sleep, Mental Health and Cognition also show common intra-domain patterns within a subtype.

Our predictions of the future subtype of individual patients in the test sample based on their baseline data, shows good accuracy in predicting disease subtypes four years later (72%

for clinical data and 67% for clinical+genetic data). 4 years is a significant time-scale for PD, which has large subtype variability. Our prediction accuracies 4 years in advance are very promising in the field of PD medicine. The explanation for the reduction in predictive accuracy with addition of genetic data may be due to: 1) the inclusion of a very limited number of genetic risk loci, 2) that genetic data isn't predictive of PD subtype within the 4-year time frame of our data or 3) that the genetic data has a large variance in the population, thus requiring a larger dataset for long-term prediction (the larger number of subtypes found by our method may indicate this). Nonetheless, from a methodological perspective, this exploratory work successfully demonstrates the inclusion of genetic data. Other biomarkers (i.e. serologic and cerebrospinal fluid biomarkers) can also be easily integrated into our analysis. Our algorithm is likely to benefit from more extensive datasets with larger populations.

A number of studies have identified PD subtypes based on baseline characteristics [7, 28–30]. In contrast to that, our innovative algorithm uses longitudinal data (or the trajectory of the different variables over time) to identify disease subtypes. In other words, our method accounts for both disease variable values as well as their progression patterns. To our knowledge, this is a novel approach. The baseline features of individual patients in a test cohort were then used to predict their future disease trajectory (prognosis). Our study represents an innovative network-based data-driven approach, that has advantages over previous methods by taking full advantage of large heterogenous, longitudinal datasets.

Despite the fact that genetic factors likely play a major role in determining PD subtypes [25], few data-driven algorithms for suptype identification exist that incorporate genetic data. Two recent studies have developed models of PD progression based on clinical, demographic and genetic data at baseline, using hierarchical cluster analysis and a Bayesian multivariate predictive inference platform, respectively, to identify PD subtypes that show significant differences in their rate of progression over time [7, 26]. Even though both of these studies thoroughly evaluate the differences of baseline subtypes in terms of long-term outcome, neither of them determines the prediction accuracy of their baseline subtype classification by repeating the subtype classification algorithm at the last time point of the follow-up period. The authors of one of the two studies [26] used the coefficient of determination R^2 as a measure of overall explanatory power of their model and found it to be 41% in the study cohort, and 9% in an independent validation cohort. However, this is a measure of how well the baseline data explain the variability at follow-up when applying their model, rather than a metric of the accuracy of subtype prediction that we introduced in our study as a novel and, in our opinion, critically important quality metric that may serve as reference when comparing our results with future subtype classification algorithms.

Our trajectory clustering method works with various types of data including clinician- and patient-reported outcome measures, genetics, physical performance measures, as well as diverse results from diagnostic investigations. This analysis uses demographics, clinician- and patient-reported data, and genetic data. In our analysis, each genetic SNP (if considered) and clinical variable is treated independently and allotted the same weight. Our algorithm allows for variable weightings, where each domain and SNP is assigned a chosen weight. However, this raises the question of how the weighting would be decided. For example, if we had allotted equal weights to one hundred SNPs in our analysis in addition to the 18 clinical variables, the genetic information would dominate the algorithm, and affect the resulting communities. On the other hand, different weighting strategies may be preferable based on the study aims. For example, if the main objective is to identify disease subtypes based on motor vs. cognitive function, one could allot equal cumulative weight to the motor and cognitive domains.

A strength of our algorithm, which is also a caveat, is that it is entirely data-driven. The level of severity of each variable relative to the baseline median is used to normalize all

variables, as opposed to the absolute value of the variable. This is done to readily compare changes in different variables. A notable example is the clinical variable, SEADL (a disability scale). SEADL is a relatively insensitive scale in the early years of PD since there is little functional disability in the years following diagnosis. Yet, in our analysis SEADL shows high progression (darker shade in later years) in Figs 3 and 5(a). It is important to note that this dark shade isn't indicative of the absolute severity. It only tells us that a larger fraction of the population in the later years has SEADL values above the baseline median of the total training population (which may be low to begin with). Like the results from any data-driven approach to identify disease subtypes, our results should be applied in practice in conjunction with medical expertise. An additional limitation of our approach is the fact that a number of choices had to be made by our team of data scientists and clinicians to create this algorithm including thresholds, the weighting scheme for all variables in the network, and variables to include. While our data-driven method is primarily agnostic, these choices are inevitably somewhat arbitrary in nature, and will have an impact on the result of the analysis. Furthermore, they may be different for different applications/datasets. Lastly, like any data-driven method, the robustness of the method is proportional to the quantity of data. Hence, while this method is suitable for heterogenous datasets such as the PPMI data, there may be other datasets that have large gaps in data collection, inconsistent times of acquiring data, too much variation in data or simply too little data- therefore, our method may not be suitable for all clinical datasets.

Our approach is innovative, adaptable, and clinically relevant. PD subtyping [31] is an area of active research but there are currently no clinically prognostic analyses in use for the management of PD. Application of an approach like ours for subtype identification as a predictive model of PD progression will help the neurologist improve clinical management of individual patients. For example, such an approach may prompt the clinician to pursue earlier, more aggressive management for those patients for whom the algorithm predicts a more rapid disease progression (i.e., 'precision medicine'). It may also guide the neurologist to perform targeted investigations (e.g. cognitive testing) in individuals based on their subtype. Finally, prediction of disease progression will improve prognostic counseling, a problem commonly encountered by clinicians, by bringing to attention disease features that are predicted to develop over the course of the disease. A natural extension of this work will be to implement this method for datasets in other chronic medical conditions. Other promising future directions include extending the TPC algorithm to incorporate and compare other network clustering approaches, such as multi-layer network clustering [8]; studying the effect of treatment on progression of disease variables, and predicting modifications of algorithm-identified subtypes as a consequence of different treatments.

A statistical analyses

We conducted statistical tests to validate our approach and demonstrate some of the differences between the subtypes identified by our TPC algorithm. These tests were focused on differentiating between the 3 subtypes illustrated in Fig 3: mixed, mild, and severe.

For each pair of subtypes, for each non-binary variable, we conducted a Mann-Whitney U test (implemented in SciPy [32]), which allows 2 groups to be compared without assuming that values are normally distributed. This test was chosen because values for many of the variables violated the normality assumption. For gender, which was binary in our analysis, we performed Fisher's exact test to calculate two-sided p-values. Fig 6 reports unadjusted p-values from these tests applied at two different time points: the baseline and the final timepoint (year 4). To achieve an overall significance level of $\alpha = 0.05$, we used the conservative Bonferroni correction to account for Type I errors due to multiple comparisons, giving us an adjusted

	Total Population	Mixed Subtype	Mild Subtype	Severe Subtype	Mixed-Mild Comparison	Mixed-Severe Comparison	Mild-Severe Comparison
		Median Values				Mann-Whitney U Test (unadjusted)	
Cognitive							
JOLO *	14.000	14.000	14.000	14.000	9.79474E-02	3.93453E-01	5.38126E-02
SDM *	42.000	43.000	46.000	37.000	4.51391E-02	1.00948E-03	**2.17537E-08**
SFT *	47.500	54.000	51.500	42.000	2.56642E-01	**5.90935E-06**	**9.67435E-07**
HVLT *	0.900	0.917	0.917	0.857	4.16284E-01	1.19052E-01	4.01845E-02
LNS *	11.000	11.000	12.000	9.000	3.37278E-01	**3.0738E-04**	**1.66226E-06**
MOCA *	27.000	28.000	28.000	26.000	4.02757E-01	6.53549E-04	**1.87142E-04**
Other							
SEADL *	95.000	95.000	95.000	90.000	3.37239E-02	3.83466E-02	**2.21909E-05**
RBDQ	4.000	6.000	3.000	5.000	**4.75543E-08**	1.97312E-01	**4.71059E-06**
ESS	6.000	7.000	4.000	7.000	**2.13653E-04**	1.46197E-01	1.29032E-03
SCOPA-AUT	9.000	11.000	5.000	12.000	**1.1846E-07**	1.30012E-01	**1.28124E-11**
GDS	2.000	3.000	1.000	2.000	**7.16102E-07**	7.71462E-03	**5.62609E-05**
STAI	65.000	76.000	60.000	65.000	**2.13091E-07**	1.08602E-03	2.97056E-02
General PD							
UPDRS1	5.500	8.000	3.000	6.000	**1.57388E-10**	3.09748E-03	**1.48489E-08**
UPDRS2	5.000	6.000	3.000	7.000	**3.91296E-04**	2.5907E-01	**4.49743E-08**
UPDRS3	20.000	19.000	17.000	23.000	2.41458E-02	1.16917E-01	**5.76769E-05**
T-UPDRS	32.000	38.000	25.000	38.000	**3.28002E-07**	3.72004E-01	**1.45972E-09**
						(A) Clinical Variables at Baseline	

	Total Population	Mixed Subtype	Mild Subtype	Severe Subtype	Mixed-Mild Comparison	Mixed-Severe Comparison	Mild-Severe Comparison
Cognitive							
JOLO *	14.000	14.000	14.000	13.000	3.70071E-01	1.28372E-01	3.81919E-02
SDM *	39.000	39.000	44.000	30.000	7.52701E-03	**3.20957E-06**	**4.06453E-14**
SFT *	46.000	54.000	52.000	39.000	3.48395E-01	**2.91609E-07**	**2.09833E-09**
HVLT *	0.913	1.000	1.000	0.750	8.30811E-02	**2.1057E-05**	**1.14352E-04**
LNS *	10.000	12.000	11.000	9.000	4.51269E-01	**5.21425E-06**	**8.53034E-09**
MOCA *	27.000	28.000	29.000	24.000	1.03701E-01	**7.2164E-07**	**2.36389E-12**
Other							
SEADL *	90.000	80.000	90.000	80.000	**3.39231E-07**	1.44041E-01	**2.76118E-10**
RBDQ	4.000	7.000	3.000	5.000	**2.34179E-07**	1.89244E-02	**1.48618E-04**
ESS	7.000	9.000	5.500	9.000	**4.83838E-05**	8.33317E-02	5.90884E-04
SCOPA-AUT	13.000	18.000	10.000	16.000	**5.16473E-07**	2.74385E-01	**3.6953E-08**
GDS	2.000	2.000	1.000	3.000	**6.27766E-08**	4.14027E-01	**7.0907E-10**
STAI	64.000	76.000	52.500	73.000	**2.89384E-10**	6.26615E-02	**2.62203E-08**
General PD							
UPDRS1	8.500	12.000	5.500	11.000	**4.72273E-10**	6.59234E-02	**5.58839E-10**
UPDRS2	10.000	13.000	5.000	12.000	**7.73458E-09**	3.30451E-01	**4.66129E-11**
UPDRS3	31.500	36.000	23.000	37.000	**4.93264E-06**	4.22977E-01	**5.67969E-08**
T-UPDRS	47.000	60.000	34.000	60.000	**3.50751E-10**	2.94173E-01	**4.74786E-12**
						(B) Clinical Variables at Year 4	

	Total Population	Mixed Subtype	Mild Subtype	Severe Subtype	Mixed-Mild Comparison	Mixed-Severe Comparison	Mild-Severe Comparison
Age	69.000	61.000	65.000	74.000	8.52716E-02	**2.6445E-09**	**5.38204E-10**
Gender	0.683	0.387	0.583	0.918	0.08278	**1.09527E-07**	**1.83353E-05**
		Fraction Male				Fisher's Exact Test	
						(C) Demographic Variables	

Fig 6. Statistical analysis. Statistical analysis comparing the 3 subtypes described in the main text: mixed, mild, and severe. Features of the total population are also listed. Medians are calculated from the raw data. Variables with negative directions are denoted by an asterisk (*). Comparisons meeting our criteria for statistical significance are shown in bold blue text. The top box (A) provides statistics for the baseline clinical variables, the middle box (B) for the year 4 clinical variables, and the bottom box (C) for demographics.

significance level of $\alpha_{\mathrm{adjusted}} = 0.05/n_c \approx 5e - 4$, where $n_c = n_p(V_d + 2V_e) = 102$ is the total number of comparisons, $n_p = 3$ is the number of pairs of subtypes, $V_d = 2$ is the number of demographic variables (age and gender), and $V_e = 16$ is the number of potentially evolving clinical variables that we assess at each of the 2 timepoints. Comparisons meeting our adjusted significance criteria are highlighted in bold in the figure.

For the baseline clinical values, we see several significant statistical differences between subtypes that support the descriptions presented in the main text. For example, compared with both the mixed and severe subtypes, the mild subtype shows significantly lower general PD severity according to 3 of 4 of the PD scales (UPDRS1, UPDRS2, and TUPDRS) as well as lower impairment according to several other scales, including the GDS (depression), SCOPA (autonomic) and RBDQ (sleep) scales. Compared with patients in the mild subtype, patients in the severe subtype show significantly greater impairment in 4 of 6 cognitive scales: SDM, SFT, LNS, and MOCA, as well as in the SEADL (disability) scale. In addition, compared even to patients in the mixed subtype, patients in the severe subtype show significantly greater impairment in the SFT and LNS cognitive scales.

For year 4 clinical values, we see even more significant differences between the subtypes than at baseline. For example, compared with the mixed and severe subtypes, the mild subtype shows significantly less impairment for all 4 PD severity scales, up from 3 at baseline. Further, the severe subtype is significantly more impaired than *both* the mixed and mild subtypes according to 5 of the cognitive scales, compared to just 2 at baseline.

We also see a significant difference in demographic variables between the subtypes. Compared to the other 2 subtypes, the severe subtype has older patients and a greater fraction of males.

While these statistical tests only compare the subtypes at two timepoints, they serve to validate our approach by highlighting some of the significant differences.

Author Contributions

Conceptualization: Sanjukta Krishnagopal, Michelle Girvan.

Data curation: Rainer von Coelln.

Formal analysis: Sanjukta Krishnagopal, Michelle Girvan.

Funding acquisition: Michelle Girvan.

Investigation: Sanjukta Krishnagopal, Rainer von Coelln, Lisa M. Shulman, Michelle Girvan.

Methodology: Sanjukta Krishnagopal.

Supervision: Rainer von Coelln, Lisa M. Shulman, Michelle Girvan.

Validation: Rainer von Coelln, Lisa M. Shulman.

Visualization: Sanjukta Krishnagopal, Michelle Girvan.

Writing – original draft: Sanjukta Krishnagopal.

Writing – review & editing: Rainer von Coelln, Lisa M. Shulman, Michelle Girvan.

References

1. Pringsheim T, Jette N, Frolkis A, Steeves TD. The prevalence of Parkinson's disease: A systematic review and meta-analysis. Movement disorders. 2014; 29(13):1583–1590. https://doi.org/10.1002/mds.25945 PMID: 24976103

2. Lees AJ, Hardy J, Revesz T. Parkinson's Disease. Lancet. 2009; 374 (9691).

3. Foltynie T, Brayne C, Barker RA. The heterogeneity of idiopathic Parkinson's disease. Journal of neurology. 2002; 249(2):138–145. https://doi.org/10.1007/PL00007856 PMID: 11985378

4. von Coelln R, Shulman LM. Clinical subtypes and genetic heterogeneity: of lumping and splitting in Parkinson disease. Current opinion in neurology. 2016; 29(6):727–734. https://doi.org/10.1097/WCO.0000000000000384 PMID: 27749396

5. Marras C, and Lang A. Parkinson's disease subtypes: lost in translation? J Neurol Neurosurg Psychiatry. 2013; 84(4):209–415. https://doi.org/10.1136/jnnp-2012-303455

6. Thenganatt MA, and Jankovic J. Parkinson disease subtypes JAMA neurology. 2014; 71(4):499–504. https://doi.org/10.1001/jamaneurol.2013.6233 PMID: 24514863

7. Fereshtehnejad SM, Zeighami Y, Dagher A, Postuma RB. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. Brain. 2017; 140(7):1959–1976. https://doi.org/10.1093/brain/awx118 PMID: 28549077

8. Fereshtehnejad SM, Postuma RB. Subtypes of Parkinson's disease: what do they tell us about disease progression? Current neurology and neuroscience reports. 2017; 17(4):34. https://doi.org/10.1007/s11910-017-0738-x PMID: 28324303

9. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature reviews genetics. 2011; 12(1):56. https://doi.org/10.1038/nrg2918 PMID: 21164525

10. Hidalgo CA, Blumm N, Barabási AL, Christakis NA. A dynamic network approach for the study of human phenotypes. PLoS computational biology. 2009; 5(4):e1000353. https://doi.org/10.1371/journal.pcbi.1000353 PMID: 19360091

11. Ivanov PC, Liu KK, Bartsch RP. Focus on the emerging new fields of network physiology and network medicine. New journal of physics. 2016; 18(10):100201. https://doi.org/10.1088/1367-2630/18/10/100201 PMID: 30881198

12. Chan SY, Loscalzo J. The emerging paradigm of network medicine in the study of human disease. Circulation research. 2012; 111(3):359–374. https://doi.org/10.1161/CIRCRESAHA.111.258541 PMID: 22821909

13. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proceedings of the National Academy of Sciences. 2007; 104(21):8685–8690. https://doi.org/10.1073/pnas.0701361104

14. Braun P, Rietman E, Vidal M. Networking metabolites and diseases. Proceedings of the National Academy of Sciences. 2008; 105(29):9849–9850. https://doi.org/10.1073/pnas.0805644105

15. Vinayagam A, Gibson TE, Lee HJ, Yilmazel B, Roesel C, Hu Y, et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. Proceedings of the National Academy of Sciences. 2016; 113(18):4976–4981. https://doi.org/10.1073/pnas.1603992113

16. Santiago JA, Potashkin JA. A network approach to clinical intervention in neurodegenerative diseases. Trends in molecular medicine. 2014; 20(12):694–703. https://doi.org/10.1016/j.molmed.2014.10.002 PMID: 25455073

17. Lewis S, Foltynie T, Blackwell AD, Robbins TW, Owen AM, Barker RA. Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. Journal of Neurology, Neurosurgery & Psychiatry. 2005; 76(3):343–348. https://doi.org/10.1136/jnnp.2003.033530

18. Castaldi PJ, Dy J, Ross J, Chang Y, Washko GR, Curran-Everett D, et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. Thorax. 2014; 69(5):416–423. https://doi.org/10.1136/thoraxjnl-2013-203601

19. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, et al. 23andMe Research Team; System Genomics of Parkinson's Disease Consortium; International Parkinson's Disease Genomics Consortium. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 2019; 18 (12):1091–1102. Genetic risk of Parkinson disease and progression: An analysis of 13 longitudinal cohorts. Neurology Genetics. 2019;5(4):e348. https://doi.org/10.1016/S1474-4422(19)30320-5 PMID: 31701892

20. Davis AA, Andruska KM, Benitez BA, Racette BA, Perlmutter JS, Cruchaga C. Variants in GBA, SNCA, and MAPT influence Parkinson disease risk, age at onset, and progression. Neurobiology of aging. 2016; 37:209–e1. https://doi.org/10.1016/j.neurobiolaging.2015.09.014 PMID: 26601739

21. Iwaki H, Blauwendraat C, Leonard HL, Liu G, Maple-Grødem J, Corvol JC, et al. Genetic risk of Parkinson disease and progression: An analysis of 13 longitudinal cohorts. Neurology Genetics. 2019; 5(4): e348. https://doi.org/10.1212/NXG.0000000000000348 PMID: 31404238

22. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment. 2008; 2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

23. Newman ME, Girvan M. Finding and evaluating community structure in networks. Physical review E. 2004; 69(2):026113. https://doi.org/10.1103/PhysRevE.69.026113

24. Javed M.A, Younis M.S, Latif S, Qadir J, Baig A. Community detection in networks: A multidisciplinary review. Journal of Network and Computer Applications. 2018; 108:87–111. https://doi.org/10.1016/j.jnca.2018.02.011

25. Cooper CA, Jain N, Gallagher MD, Weintraub D, Xie SX, Berlyand Y, et al. Common variant rs356182 near SNCA defines a Parkinson's disease endophenotype. Annals of clinical and translational neurology. 2017; 4(1):15–25. https://doi.org/10.1002/acn3.371 PMID: 28078311

26. Latourelle JC, Beste MT, Hadzi TC, Miller RE, Oppenheim JN, Valko MP, et al. Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed Parkinson's disease: a longitudinal cohort study and validation. The Lancet Neurology. 2017; 16(11):908–916. https://doi.org/10.1016/S1474-4422(17)30328-9 PMID: 28958801

27. Iwaki H, Blauwendraat C, Leonard H, Kim J, Liu G, Maple-Grødem J, et al. Genomewide association study of Parkinson's disease clinical biomarkers in 12 longitudinal patients' cohorts. Movement Disorders. 2019; 3434(12):1839. https://doi.org/10.1002/mds.27845

28. Fereshtehnejad SM, Romenets SR, Anang JB, Latreille V, Gagnon JF, Postuma RB. New clinical subtypes of Parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes. JAMA neurology. 2015; 72(8):863–873. https://doi.org/10.1001/jamaneurol.2015.0703 PMID: 26076039

29. Lawton M, Ben-Shlomo Y, May MT, Baig F, Barber TR, Klein JC, et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. J Neurol Neurosurg Psychiatry. 2018; 89(12):1279–1287. https://doi.org/10.1136/jnnp-2018-318337 PMID: 30464029

30. Rajput AH, Rajput ML, Ferguson LW, Rajput A. Baseline motor findings and Parkinson disease prognostic subtypes. Neurology. 2017; 89(2):138–143. https://doi.org/10.1212/WNL.0000000000004078 PMID: 28592451

31. Kivela M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. Journal of complex networks. 2014; 2(3):203–271. https://doi.org/10.1093/comnet/cnu016

32. https://docs.scipy.org/doc/scipy-0.19.1/reference/generated/scipy.stats.mannwhitneyu.html