frontiers | Frontiers in Aging Neuroscience

TYPE Original Research
PUBLISHED 21 February 2023
DOI 10.3389/fnagi.2023.1047017

OPEN ACCESS

EDITED BY
Jinghong Ma,
Xuanwu Hospital, Capital Medical University, China

REVIEWED BY
Shady Rahayel,
McGill University Health Centre, Canada
Stephen Smith,
University of York, United Kingdom

*CORRESPONDENCE
Junwu Tu
✉ tujw@shanghaitech.edu.cn

SPECIALTY SECTION
This article was submitted to Alzheimer's Disease and Related Dementias, a section of the journal Frontiers in Aging Neuroscience

RECEIVED 17 September 2022
ACCEPTED 02 February 2023
PUBLISHED 21 February 2023

CITATION
Ma L-Y, Feng T, He C, Li M, Ren K and Tu J (2023) A progression analysis of motor features in Parkinson's disease based on the mapper algorithm. *Front. Aging Neurosci.* 15:1047017. doi: 10.3389/fnagi.2023.1047017

A progression analysis of motor features in Parkinson's disease based on the mapper algorithm

Ling-Yan Ma[1,2], Tao Feng[1,2,3], Chengzhang He[4], Mujing Li[4], Kang Ren[5,6] and Junwu Tu[4]*

[1]Department of Neurology, Center for Movement Disorders, Beijing Tiantan Hospital, Capital Medical University, Beijing, China, [2]Department of Neurology, China National Clinical Research Center for Neurological Disease, Beijing, China, [3]Parkinson's Disease Center, Beijing Institute for Brain Disorders, Beijing, China, [4]Institute of Mathematical Sciences, ShanghaiTech University, Shanghai, China, [5]GYENNO Science Co., LTD., Shenzhen, China, [6]Department of Neurology, HUST-GYENNO Central Neural System Intelligent Digital Medicine Technology Center, Wuhan, China

Background: Parkinson's disease (PD) is a neurodegenerative disease with a broad spectrum of motor and non-motor symptoms. The great heterogeneity of clinical symptoms, biomarkers, and neuroimaging and lack of reliable progression markers present a significant challenge in predicting disease progression and prognoses.

Methods: We propose a new approach to disease progression analysis based on the mapper algorithm, a tool from topological data analysis. In this paper, we apply this method to the data from the Parkinson's Progression Markers Initiative (PPMI). We then construct a Markov chain on the mapper output graphs.

Results: The resulting progression model yields a quantitative comparison of patients' disease progression under different usage of medications. We also obtain an algorithm to predict patients' UPDRS III scores.

Conclusions: By using mapper algorithm and routinely gathered clinical assessments, we developed a new dynamic models to predict the following year's motor progression in the early stage of PD. The use of this model can predict motor evaluations at the individual level, assisting clinicians to adjust intervention strategy for each patient and identifying at-risk patients for future disease-modifying therapy clinical trials.

KEYWORDS

progression analysis, Parkinson's disease, mapper algorithm, Markov chain, prediction model

1. Introduction

Parkinson's disease is a neurodegenerative disease with a broad spectrum of motor symptoms including bradykinesia, rigidity, resting tremor, and postural and gait impairments (Selikhova et al., 2009). In the clinical course of PD, both linear (Gottipati et al., 2017; Holden et al., 2018) and non-linear progression (Vu et al., 2012; Reinoso et al., 2015) have been reported in the advancement of motor and non-motor symptoms. The substantial heterogeneity in the presentation of clinical phenotypes, genetics, pathology, and disease progression (Foltynie et al., 2002; Selikhova et al., 2009; Ma et al., 2015) and lack of reliable progression markers of neurodegeneration present a major challenge for prediction of progression and accurate prognoses, hampering advances in PD trials, and the clinical routine determining therapeutic efficacy. In an era of increasing focus on individualized management and disease-modifying therapies, there is a need to develop useful tools to predict each patient's motor progression with high accuracy.

The current literature on PD progression consists largely of associative analyses and a few prognostic models. The prognostic models include logistic regression and Bayesian classification models to predict cognitive impairment (Schrag et al., 2017; Hogue et al., 2018; Gramotnev et al., 2019), machine-learning, random survival forests to predict time to initiation of symptomatic treatment (Simuni et al., 2016) and disease progression (Latourelle et al., 2017; Severson et al., 2021). Besides, partial least squares path modeling (PLS-PM), combined with MRI biomarkers, were used to predict progression subtypes and cognitive impairment in prodromal PD (Pyatigorskaya et al., 2021; Rahayel et al., 2021). Based on the Parkinson's Progression Markers Initiative (PPMI) database, we previously built five regression models to predict PD motor progression represented by the coming year's Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part III score, finding adjusted R^2 values of three different categories of regression model, linear, Bayesian, and ensemble, all reached 0.75 (Ma et al., 2020).

In this study, we propose a new approach to disease progression analysis based on topological data analysis (TDA), or the mapper algorithm to be precise. The mapper algorithm was introduced in Singh et al. (2007) by Singh-Memoli-Carlsson as a way of capturing topological/geometric informations of a point cloud dataset possibly in a high dimensional Euclidean space. Roughly speaking, it may be viewed as an algorithm to compute a given dataset's geometric "shape" by certain combinatorial object which, in the simplest form, may be a graph or a polyhedron. In the case of analyzing patients' data, the method has been successfully implemented in a variety of circumstances (see for example Nicolau et al., 2011; Li et al., 2015; Rossi-deVries et al., 2018; Dagliati et al., 2020).

It is always difficult to predict PD because of great heterogeneity, including subtypes, markers, and various scales. Only by combining clinical presentation and mathematical methods, selecting appropriate parameters and applying appropriate methods can the accuracy of prediction model be improved. Based on PPMI data and our previous predicting models, we aim to improve our multiple dynamic prediction model *via* mapper algorithm in this study. Similarly, general information and classical clinical scales, which are routinely and easily performed in clinical activities, were used to predict motor progression, displayed in the form of the MDS–UPDRS Part III score. These inexpensive and easily readily available clinical data can facilitate widespread implementation of this cost-efficient predictive model in real world applications.

2. Materials and methods

2.1. Feature selection and data pre-processing

The data were obtained from the PPMI database. The PPMI is an international, multicenter, prospective study designed to discover, and validate biomarkers of disease progression in newly diagnosed PD participants (National Clinical Trials identifier NCT01141023). Each PPMI recruitment site received approval from an institutional review board or ethics committee on human experimentation before study initiation. Written informed consent for research was obtained from all individuals participating in the study. The PPMI database was accessed on December 16, 2022, to obtain data from 943, 379, 324, 256, 268 visits for five consecutive years. For up-to-date information on the study, please visit www.ppmi-info.org.

Since the mapper algorithm is a way of computing the "shape" of a given data set in $\mathbb{R}^N$, if the dimension N is too large while the data set is relatively small, the shape would only be a collection of sparse points. Thus, our first step uses a topological method to reduce the number of features introduced in Kraft (2016). The idea behind this feature selection method is that we could eliminate a feature if it does not cause a big change in the underlying topology (calculated using persistent homology) of the data sets. We refer to the article (Kraft, 2016) for more details.

In our case, for the feature selection, we first consider the following listed 29 features mostly used in Ma et al. (2020). We have added a feature "symptom" which is given by the sum symptom1, symptom2, symptom3, and symptom4, with

- Symptom1: Initial symptom (at diagnosis)—Resting Tremor
- Symptom2: Initial symptom (at diagnosis)—Rigidity
- Symptom3: Initial symptom (at diagnosis)—Bradykinesia
- Symptom4: Initial symptom (at diagnosis)—Postural Instability

All these variables are binary such that it is 0 if No symptom or unknown; 1 if Symptom present at diagnosis.

In general, the features we consider are inexpensive and easily readily available clinical data. Each of the coordinates is normalized to $[0, 1]$. In the coordinate given by the UPDRS III score, we also performed a clamping at 0.7.[1] These features are listed as follows:

*updrs*3, *age*, *NP*1*APAT*, *scopa*, *YEAR*, *NP*1*FATG*, *moca*, *symptom*, *NP*1*ANXS*, *gds*, *PD_MED_USE*, *symptom*2, *NP*1*HALL*, *ageonset*, *NP*1*COG*, *NP*1*DPRS*, *rem*, *ess*, *symptom*1, *DOMSIDE*, *"PATNO,"* *gen*, *symptom*4, *fampd_new*, *NP*1*DDS*, *duration*, *td_ pig*, *quip*, *symptom*3.

In the above we have ordered the features according to their Pearson's correlation coefficients with the UPDRS III score. An important point to note is that, excluding the UPDRS III score itself, the maximal of these Pearson's correlation coefficients is 0.27, which shows that their correlation with the UPDRS III score is in general highly non-linear. This is an ideal context to use our topological data analysis (TDA) method as it is a tool developped to handle non-linear correlations.

Then, we use the persistent homology to reduce the number of features (Kraft, 2016). In our case, Figure 1 illustrates the persistent homology, when passing from the first eight features to seven features, has a big difference. This tells us we should stop eliminating features. The remaining 8 selected features are listed as in Table 1.

From the PPMI data, we select these features for each patient's data to form a point cloud $S^{\text{PPMI}} \subset \mathbb{R}^8$, of size $|S^{\text{PPMI}}| = 2,389$, consisting of 481 distinguished patients.

1 The reason for clamping is that very rarely we have patients with UPDRS III score > 0.7 after normalization.

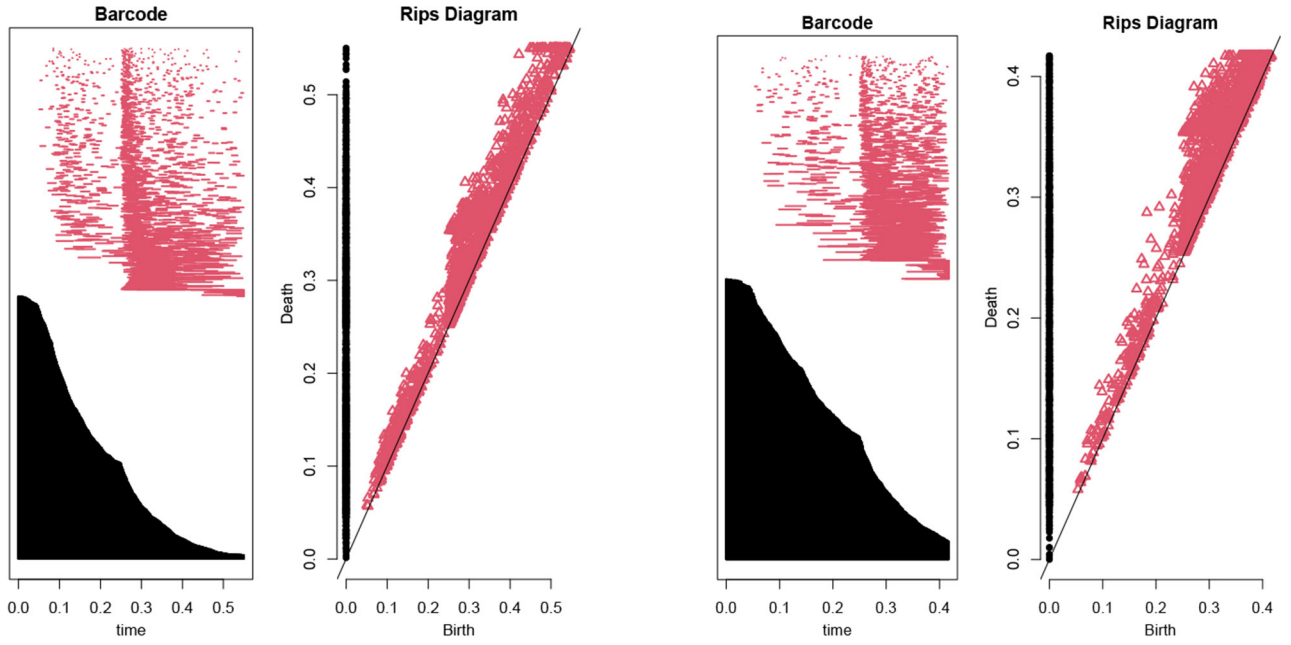

FIGURE 1
The barcodes and rips diagrams both illustrate the persistent homology (Lum et al., 2013) of a given dataset. The two pictures compare the persistent homology in the case of seven features with the case of eight features. Observe that the right hand side has considerably more persistence (in both black and red markings) compared with the left hand side.

TABLE 1 List of selected features.

Features	Meaning
Age	Age of patient when data is collected
updrs3	UPDRS III score (OFF)
np1apat	APATHY
np1fatg	FATIGUE
np1anxs	ANXIOUS MOOD
Moca	Montreal cognitive assessment (MOCA) score
Scopa	Scales for outcomes in Parkinson's disease (SCOPA)-AUT total score
Symptom	Symptom1 + Symptom2 + Symptom3 + Symptom4

2.2. The mapper algorithm

The mapper algorithm introduced by Singh-Memoli-Carlsson (Singh et al., 2007) is a method to analyze high dimensional data based on ideas from topology—a branch of mathematics to study complex shapes of geometric objects.

Roughly speaking, the mapper algorithm consists of several steps as illustrated in Figure 2.

(A) A point cloud data $S \subset \mathbb{R}^N$ whose topological/geometric properties we would like to study.
(B) A choice of d filter functions on S.

$$f = (f_1, \ldots, f_d) : S \to \mathbb{R}^d.$$

(C) Choose a covering of the image of f by boxes:

$$\mathrm{Im}(f) \subset \bigcup_{\alpha} B_\alpha$$

with each B_α a box in $\mathbb{R}^d$. Put the data S into overlapping bins by taking $S_\alpha := f^{-1}(B_\alpha)$.
(D) Cluster each bin S_α and create a simplicial complex recording the intersection pattern between the clusters. Often a truncated version is used as mapper's output: the result yields a graph whose vertices correspond to the clusters, and an edge is created whenever two clusters have non-empty intersection.

2.3. Construction of Markov chains

We shall apply the mapper algorithm to the point cloud S^{PPMI} from the previous subsection. Recall that $S^{\mathrm{PPMI}} \subset \mathbb{R}^8$ is the sample space of patients' data extracted from the raw PPMI data. Using the mapper algorithm, assume that we have obtained m clusters $C_1, \ldots, C_m$ so that $S^{\mathrm{PPMI}} = C_1 \cup \cdots \cup C_m$. Note that these clusters can possibly intersect with each other.

Let $P \subset S^{\mathrm{PPMI}} \times S^{\mathrm{PPMI}}$ be a subset. We proceed to use P to obtain a Markov chain on the set of clusters $C_1, \ldots, C_m$. For a pair of data $(x, y) \in P$, if $x \in C_i$ and $y \in C_j$, we consider it as an arrow from the cluster C_i to C_j. This yields a multi-graph (possibly with multiple edges between vertices) whose vertices are the clusters $C_1, \ldots, C_m$. Then we use informations of this multi-graph to obtain a Markov matrix. More precisely, for each pair of indices (i, j) with

10.3389/fnagi.2023.1047017

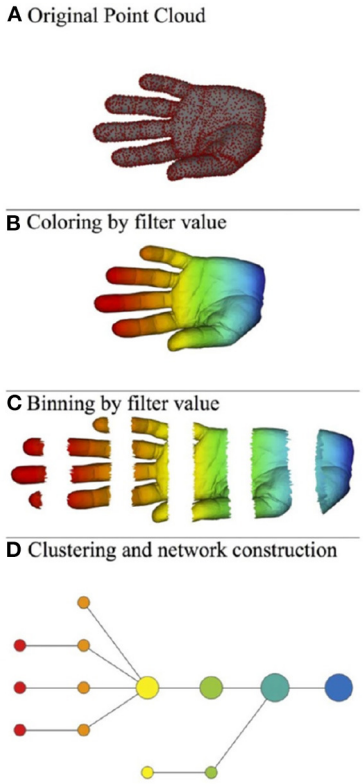

FIGURE 2
Illustration of the mapper algorithm in the case of a point cloud S in $\mathbb{R}^2$, and with the filter function given by the horizontal projection. The outcome of this algorithm is the bottom graph. **(A)** Original point cloud, **(B)** Coloring by filter value, **(C)** Binning by filter value, and **(D)** Clustering and network construction.

$1 \leq i, j \leq m$, we define

$$M_{ij}^P := \begin{cases} \frac{|\{(x,y)\in P | x\in C_i, y\in C_j\}|}{|\{(x,y)\in P | x\in C_i\}|} & \text{if } \{(x,y)\in P \mid x\in C_i\} \neq \emptyset \\ \delta_{ij} & \text{if } \{(x,y)\in P \mid x\in C_i\} = \emptyset \end{cases} \quad (1)$$

2.3.1. Computing expected growth

For each $1 \leq j \leq m$, denote by

$$E_j := \frac{1}{|C_j|} \cdot \sum_{y\in C_j} \text{updrs3}(y)$$

the expected value of the UPDRS III score of the cluster C_j.

The expected growth of a patient's UPDRS III of a fixed PD medication type i is computed as follows.

(1) Fix the medication type index $0 \leq i \leq 7$. Consider the distribution of patients with medication type i in each cluster, i.e., for each $1 \leq j \leq m$ denote by

$$d_j := |\{x \in C_j \mid \text{The data x is from a patient with PD medication type} i.\}|$$

(2) Form the initial probability vector that a type i patient belongs to each cluster:

$$w := \frac{1}{\sum_{j=1}^m d_j}(d_1, \ldots, d_m).$$

(3) The expected growth in 1 year of such a patient is then computed by

$$\Delta := \sum_{j=1}^m w_j \cdot \Delta_j$$

where $\Delta_j = \sum_{l=1}^m M_{jl}^{P_i}(E_l - E_j)$ is the expected growth of the UPDRS III score for a patient in the cluster C_j.

2.4. Prediction models

As a second application, we use the Markov chains obtained in the previous paragraph to build a prediction model for a patient's UPDRS III score in the next year. This is done in several steps:

(a) Given a patient's current year data $x \in \mathbb{R}^8$, we first produce an initial probability vector

$$v = (v_1, v_2, \ldots, v_m)$$

where recall that m is the number of clusters in the mapper output. See Equation (2) for the definition of v. In other words, v_j is the probability of x lie inside the j-th cluster C_j.

(b) Then compute the action of the Markov chain on the vector v to obtain

$$v^\dagger := v \cdot M^P = (v_1^\dagger, \ldots, v_m^\dagger)$$

(c) The predicted UPDRS III score is then equal to

$$p(x) := \sum_{j=1}^m v_j^\dagger \cdot E_j$$

where as before $E_j := \frac{1}{|C_j|} \cdot \sum_{y\in C_j} \text{updrs3}(y)$ is the expected value of the UPDRS III score of the cluster C_j.

The first step (a) needs more explanation, and is realized as follows. Fix a positive integer $\mu > 0$, and a positive real number $\sigma > 0$. We find the first μ nearest point $a_1, \ldots a_\mu \in S^{\text{PPMI}}$ to the given point x. Then use the equation

$$c \cdot \sum_{k=1}^{\mu} e^{-\frac{||x-a_k||^2}{\sigma^2}} = 1$$

to determine a constant c. For each $1 \leq k \leq \mu$, the point a_k may belong to several clusters. Denote its multiplicity by

$$l_k := |\{1 \leq i \leq m \mid a_k \in C_i\}|$$

At this point, it is tempted to set the initial probability vector by formula

$$v_i = \sum_{1\leq k\leq \mu,\ a_k\in C_i} \frac{1}{l_k} \cdot c \cdot e^{-\frac{||x-a_k||^2}{\sigma^2}}.$$

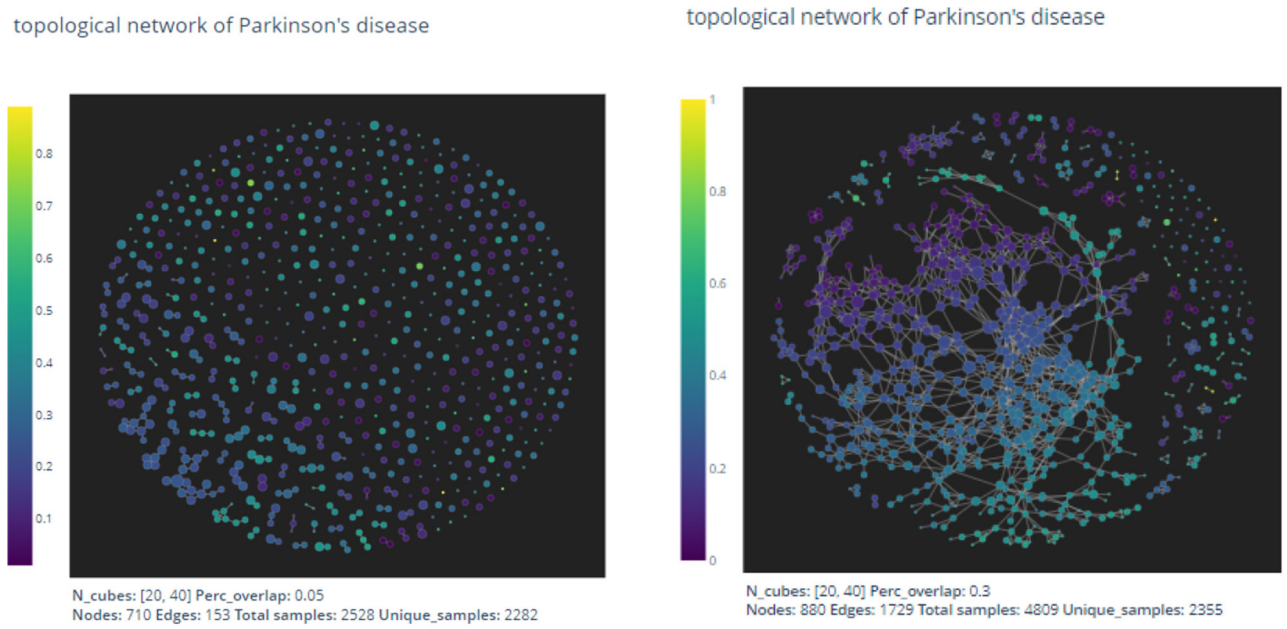

FIGURE 3
The pictures are out-puts of the mapper algorithm with parameters given by $n_1 = 20$, $n_2 = 40$, $p = 0.05$ and $n_1 = 20$, $n_2 = 40$, $p = 0.3$, respectively.

However, observe that already in the definition of M^P (see Equation 1), it is possible that a cluster C_i is not the source of any arrows, i.e.,

$$\{(x, y) \in P \mid x \in C_i\} = \emptyset.$$

In this case, it is not possible to use such type of clusters to make predictions for the next year's data. Thus, we set the initial probability at such a cluster by zero, and rescale the resulting vector by a constant to obtain the desired initial probability vector. Explicitly, we set the initial probability vector $v = (v_1, v_2, \ldots, v_m)$ by

$$v_i := \begin{cases} \text{const} \cdot \sum_{1 \le k \le \mu,\ a_k \in C_i} \frac{1}{l_k} \cdot c \cdot e^{-\frac{\|x - a_k\|^2}{\sigma^2}} & \text{if } \{(x, y) \in P \mid x \in C_i\} \neq \emptyset \\ 0 & \text{if } \{(x, y) \in P \mid x \in C_i\} = \emptyset \end{cases} \quad (2)$$

In this paper, we shall fix the parameters to be $\mu = 14$ and $\sigma = 0.0378$.

3. Results

3.1. Mapper outputs

We apply the Kepler mapper program 1.4.1 (van Veen et al., 2019a,b) to the point cloud set S^{PPMI} with a 2-dimensional filter function

$$f = (\text{age}, \text{updrs3}) : S \to \mathbb{R}^2$$

given by two coordinate projections in the direction of "age" and "updrs3." The output graph is shown in Figure 3.

As expected by the formation of the mapper algorithm, larger percentage of overlaps naturally leads to more non-empty intersections between clusters, and hence the graph on the right appears to have more edges than the left one.

In the two dimensional mapper algorithm, there are three parameters to choose:

- n_1: Number of intervals in the "updrs3" direction.
- n_2: Number of intervals in the "age" direction.
- p: Percentage of overlaps in both direction.

There exists no general method to determine appropriate parameters in the mapper algorithm. In the next section, we shall use the mapper output to construct a prediction model for the UPDRS III scores of patients. We then use the precision value of the resulting prediction model to evaluate and thus optimize the parameters.

3.2. Markov chains

From the PPMI data, there are eight different types of patients according to their usage of PD medications, as shown in Table 2.

Denote by $P_i \subset S^{\text{PPMI}} \times S^{\text{PPMI}}$, $0 \le i \le 7$ the subset consisting of pairings (x, y) such that the data x and y are two consecutive years' data from the same patient (i.e., a progression by 1 year), and that the patient's usage of PD medication is of type i in the above table. For $i = 0$ and $i = 1$ we have depicted the corresponding two Markov chains in Figure 4 (with mapper parameters set to be $n_1 = 20, n_2 = 40, p = 0.05$).

3.3. PD medication type analysis

As a first application of the Markov chains M^{P_i} obtained from the previous paragraph. We use it to compute the expected growth of a patient's UPDRS III score according to the patient's PD medication type. The computed results are shown in Table 3.

The expected growth of PD patients with a particular type of medication certainly may depend on the particular choice of medication to begin with. Thus, it makes sense to perform an un-biased comparison with what happens if the medication type $i \neq 0$ group of patients were not given any medication. To do this, consider the following probability distribution $(p_1, \ldots, p_m)$ on the set of clusters defined by

$$p_j := \frac{\text{type } i \text{ patients in } C_j}{\text{all type } i \text{ patients}}.$$

We can calculate the expected growth viewed as un-medicated patients under same distribution using the Markov chain $M_{jk}^{P_0}$:

$$\Delta_i' := \sum_j \sum_k p_j M_{jk}^{P_0} (E_k - E_j)$$

TABLE 2 List of medication types.

Type index	Medication
0	Unmedicated
1	Levodopa
2	Dopamine agonist
3	Other
4	Levodopa + other
5	Levodopa + dopamine agonist
6	Dopamine agonist + other
7	Levodopa + dopamine agonist + other

The difference between Δ_i' and the actual expected growth Δ_i would measure the benefit of the i-th type medication to reduce the growth of patients' UPDRS III scores. Calculations demonstrate solid medication effects in the cases of type 4, 5, and 6, as shown in Table 4. Observe that patients in medication type 5 and 6 have relatively small expected growth of UPDRS score in Table 3. The un-biased analysis gives at least a partial explanation for this: for these two groups of patients medication effects are rather significant.

TABLE 3 Expected growth of UPDRS III score associated with different medication types.

PD medication type index	Expected growth of UPDRS III score
0	2.17
1	2.24
2	2.51
3	3.37
4	2.19
5	0.30
6	0.88
7	1.85

TABLE 4 Un-biased medication effects in medication type 4, 5, and 6.

Medication type i	Δ_i	Δ_i'	$\Delta_i' - \Delta_i$
4	2.19	2.50	0.31
5	0.30	1.30	1
6	0.88	2.29	1.41

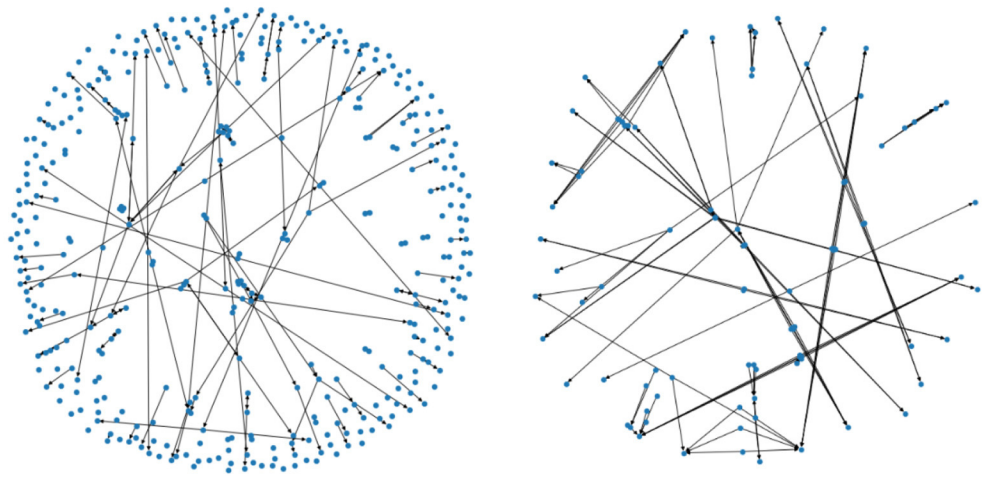

FIGURE 4
The two figures illustrate two Markov chains associated with medication type 0 and 1, respectively. Its nodes are derived from the outputs of the mapper algorithm.

TABLE 5 Statistics of the TDA method.

Medication	R^2 score	MAE	MSE	Max error	Hit in percentage %
0	0.67	−0.0121	0.00597	0.216	62.6
1	0.726	−0.00807	0.00607	0.222	82.7
2	0.966	−0.00542	0.000396	0.0127	97.0
3	0.872	−0.000186	0.0015	0.14	88.4
4	0.642	−0.0142	0.00691	0.247	77.6
5	0.749	−0.012	0.00515	0.217	87.7
6	0.499	−0.0023	0.00698	0.274	78.6
7	0.953	0.000792	0.00073	0.0663	94.3

TABLE 6 Comparison between statistics of the TDA method with classical regression methods.

	R^2 score	MAE	MSE	Max error	Hit in percentage %
TDA	0.726	−0.00807	0.00607	0.222	82.7
Linear regression	0.607	0.0632	0.00704	0.384	55.7
Ridge regression	0.642	0.0693	0.00794	0.315	46.8
Bayesian regression	0.689	0.0635	0.0069	0.283	55.5
Random forest	0.733	0.0372	0.00593	0.562	78.6
Gradient boosting	0.783	0.0364	0.00481	0.311	79.2

3.4. Statistics of the prediction models

To test the validity of our prediction model described above, for each PD medication type index $0 \leq i \leq 7$, we perform a statistical study of its accuracy as follows.

(1) First take out a point $(x_0, y_0) \in P_i \subset S^{\text{PPMI}} \times S^{\text{PPMI}}$, run the prediction model with $S - \{x_0\}$ to obtain the predicted next year's UPDRS III score $p(x_0)$.
(2) Do step (1) for all points (x_0, y_0) in P_i. Then perform a statistical study between the predicted score $p(x_0)$ with the actual next year's score y_0.

Table 5 shows the statistics of our prediction models in each PD medication type. The R^2 score, MAE, MSE and Max Error are well-known statistical measures. We explain the last column "hit percentage." In the evaluation of UPDRS III score (a total of 132 points), medical experiences usually permits a variation of ± 5 points. In our data set S^{PPMI}, the difference between maximal score and the minimal score is 80. Since we have normalized this score to $[0, 1]$, a variation of ± 5 absolute points would corresponds to ± 0.0625 after normalization. The "hit-in percentage" is the percentage of the prediction score $p(x_0)$ "hit-in" the interval $[y_0 - 0.0625, y_0 + 0.0625]$ since we regard such a prediction as being a successful one.

3.5. Comparison with classical regression methods

The statistics shown above should be compared with an earlier prediction model (Ma et al., 2020). In *loc. cit.* the authors used classical methods such as Linear Regression, Bayesian Regression, and so on. For example, in the case of P_1, the comparison of statistics of our TDA method with classical methods is shown in Table 6.

This shows that the mapper algorithm combined with Markov chain construction is more efficient than the more classical regression methods in the study of progression analysis of Parkinson's disease.

4. Discussion

In this study, we develop a new predictive model for motor progression in patients with early PD by mapper algorithm, which we report 62.5% accuracy in the group of un-medicated patients (Medication type 0); while in other medication types, the accuracy increased, fluctuating between 77.6 and 97% (Medication type 1–7). Also, we compared different methods in the analysis of PD progression and found that mapper algorithm combined with Markov chain construction is more efficient than the more classical regression methods. This prediction model is an upgrade of our previous prediction model, which improves the accuracy and has better stability. Our findings indicate that the models can practically predict the MDS-UPDRS Part III score of the coming year based on the clinically available characteristics obtained in the current year.

There are a growing number of clinical prediction models of the progression of PD, which vary from the choices of predictive values according to different objectives. Latourelle et al. developed and validated a comprehensive multivariable prognostic model based on the PPMI database (Latourelle et al., 2017). In this model, they obtained a R^2 of 41% in PPMI database and 9% in LABS-PD database that used for external validation. This reduction of

R^2 could be offset by increasing the sample size. As in Lu et al. they developed a progression model based on the videos of MDS-UPDRS tests to estimate the motor severity of PD, in which they obtained a classification accuracy of 72% and F1-score of 0.51 (Lu et al., 2021).

Eight variables were enrolled in this model, including age, MDS-UPDRS III, NP1 apathy score, NP1 fatigue score, NP1 anxiety score, MOCA, SCOPA-AUT, and initial symptoms. These variables contain quantification of motor (MDS-UPDRS III) and non-motor symptoms (apathy, cognitive dysfunction, fatigue and anxiety), all of which contribute to the progression of PD.

Previous studies have identified that cognitive impairment at baseline is correlated with faster disease progression and greater motor impairment (Velseboer et al., 2013; Fereshtehnejad et al., 2015; Reinoso et al., 2015). Apart from UPDRS values, signs of cognitive decline, orthostatic hypotension and rapid eye movement sleep behavior disorder at baseline, could also suggest a much faster decline in motor symptoms. An increase in L-dopa non-responsive symptoms, which suggest a diffuse destruction of extra-nigrostriatal pathways in parallel with the nigrostriatal pathway (Velseboer et al., 2013) may in part explain the situation.

Overall, PD is a neurodegeneration disease and all the patients suffer from progressive aggravation. The expected growth of motor score varies greatly due to different medication types. The rate of progress of patients with no medication is 2.17 per year, which is representative of PD's natural course. Anti-PD drugs can improve patients' motor symptoms, while the expected growth of UPDRS III score in patients taking medicine is lower than type 1. We also found the expected growth of UPDRS III score in groups 5 (levodopa + dopamine agonist) and 6 (dopamine agonist + other) is lower than other types, indicating that dopamine agonists might improve motor dysfunction better or exist potential disease-modifying effect. However, given the complexity of drugs regulation and interactions with patients, further interpretation should be given cautiously. In addition, according to the type of medication used by the patients, the accuracy of prediction model in the patients taking the anti-PD medication was improved compared to patients with no medication, ranging from 77.6 to 97%. The reason is that in the type 0 case, patients' UPDRS III score could experience a "jumping" phenomenon, thus making our continuous topological method not as effective as in the case of other medication types. In fact, identifying features of this jumping phenomenon is itself an interesting question which we plan to further investigate in a future work.

There are also some limitations in this study. First, the variability and subjectiveness of measures of the motor and non-motor scores within the PPMI dataset may exist. Second, due to limited PD patients, only uniform predictions across subtypes were made without consideration of PD subtypes. Third, we just predict the MDS-UPDRS Part III total score in the predict model, and no subdivision prediction was made for a single item or symptom category score (such as limb rigidity, central axis slowing, tremor, gait, etc.). Finally, our analysis was based on the early stage of PD. As a result, this model cannot be apply to patients with advanced PD for motor prediction.

In this study, by using mapper algorithm, we apply relatively fewer parameters to achieve better results than the previous models, provide accuracy in the range of 62.5 − 97.0% in predicting motor progression depending on different medication types. The use of this model can predict motor evaluations at the individual level, assisting clinicians to adjust intervention strategy for each patient and identifying at-risk patients for future disease-modifying therapy clinical trials.

Data availability statement

Publicly available datasets were analyzed in this study. The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: www.ppmi-info.org.

Author contributions

L-YM: research directions throughout the process and provides medical advise for feature selection together with TF. TF: medical advise for feature selection, suggest to produce practical applications of the algorithm, and such as comparing different medication types using the algorithm. CH: algorithm implementations and coding. ML: data pre-processing. KR: current collaboration and discussions throughout different stages of the program. JT: algorithm implementations involved in the program. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by Natural Science Foundation of China (Grant Nos. 82071422 and 81571226), National Keypoint Research and Invention Program of the Thirteenth (2016YFC1306501), Capital Characteristic Clinic Project (Z171100001017041), Beijing Municipal Science and Technology Commission (Grant Nos. Z151100003915117 and Z151100-003915150), and Beijing Natural Science Foundation (No. 7164254). This work was supported by grants from the Michael J. Fox Foundation for Parkinson's Research and the National Institute of Neurological Disorders and Stroke (1P20NS092529-01). PPMI–a public-private partnership–was funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbvie, Avid, Biogen, BioLegend, Bristol-Myers Squibb, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Roche, Sanofi, Servier, Takeda, Teva, UCB, and Golub Capital.

Conflict of interest

KR was employed by GYENNO Science Co., LTD.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor JM declared a shared parent affiliation with the authors L-YM and TF at the time of review.

Publisher's note

References

Dagliati, A., Geifman, N., Peek, N., Holmes, J. H., Sacchi, L., Bellazzi, R., et al. (2020). Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records. *Artif. Intell. Med*. 108, 101930. doi: 10.1016/j.artmed.2020.101930

Fereshtehnejad, S. M., Romenets, S. R., Anang, J. B., Latreille, V., Gagnon, J. F., and Postuma, R. B. (2015). New clinical subtypes of Parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes. *JAMA Neurol*. 72, 863–873. doi: 10.1001/jamaneurol.2015.0703

Foltynie, T., Brayne, C., and Barker, R. A. (2002). The heterogeneity of idiopathic Parkinson's disease. *J Neurol*. 249, 138–145. doi: 10.1007/PL00007856

Gottipati, G., Karlsson, M. O., and Plan, E. L. (2017). Modeling a composite score in Parkinson's disease using item response theory. *AAPS J*. 19, 837–845. doi: 10.1208/s12248-017-0058-8

Gramotnev, G., Gramotnev, D. K., and Gramotnev, A. (2019). Parkinson's disease prognostic scores for progression of cognitive decline. *Sci. Rep*. 9, 17485. doi: 10.1038/s41598-019-54029-w

Hogue, O., Fernandez, H. H., and Floden, D. P. (2018). Predicting early cognitive decline in newly-diagnosed Parkinson's patients: a practical model. *Parkinsonism Relat. Disord*. 56, 70–75. doi: 10.1016/j.parkreldis.2018.06.031

Holden, S. K., Finseth, T., Sillau, S. H., and Berman, B. D. (2018). Progression of MDS-UPDRS scores over five years in *de novo* Parkinson disease from the Parkinson's progression markers initiative cohort. *Mov. Disord. Clin. Pract*. 5, 47–53. doi: 10.1002/mdc3.12553

Kraft, R. (2016). *Illustrations of data analysis using the mapper algorithm and persistent homology. TRITA-MAT-E* (Master Thesis). Royal Institute of Technology, SCI School of Engineering Sciences, Stockholm, Sweden.

Latourelle, J. C., Beste, M. T., Hadzi, T. C., Miller, R. E., Oppenheim, J. N., Valko, M. P., et al. (2017). Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed Parkinson's disease: a longitudinal cohort study and validation. *Lancet Neurol*. 16, 908–916. doi: 10.1016/S1474-4422(17)30328-9

Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., et al. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med*. 7, 311ra174. doi: 10.1126/scitranslmed.aaa9364

Lu, M., Zhao, Q., Poston, K. L., Sullivan, E. V., Pfefferbaum, A., Shahid, M., et al. (2021). Quantifying Parkinson's disease motor severity under uncertainty using MDS-UPDRS videos. *Med. Image Anal*. 73, 102179. doi: 10.1016/j.media.2021.102179

Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., et al. (2013). Extracting insights from the shape of complex data using topology. *Sci. Rep*. 3, 1236. doi: 10.1038/srep01236

Ma, L. Y., Chan, P., Gu, Z. Q., Li, F. F., and Feng, T. (2015). Heterogeneity among patients with Parkinson's disease: cluster analysis and genetic association. *J Neurol Sci*. 351, 41–45. doi: 10.1016/j.jns.2015.02.029

Ma, L. Y., Tian, Y., Pan, C. R., Chen, Z. L., Ling, Y., Ren, K., et al. (2020). Motor progression in early-stage parkinson's disease: a clinical prediction model and the role of cerebrospinal fluid biomarkers. *Front. Aging Neurosci*. 12, 627199. doi: 10.3389/fnagi.2020.627199

Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. U.S.A*. 108, 7265–7270. doi: 10.1073/pnas.1102826108

Pyatigorskaya, N., Yahia-Cherif, L., Valabregue, R., Gaurav, R., Gargouri, F., Ewenczyk, C., et al. (2021). Parkinson disease propagation using MRI biomarkers and partial least squares path modeling. *Neurology* 96, e460-e471. doi: 10.1212/WNL.0000000000011155

Rahayel, S., Postuma, R. B., Montplaisir, J., Miši,ć B., Tremblay, C., Vo, A., et al. (2021). A prodromal brain-clinical pattern of cognition in synucleinopathies. *Ann. Neurol*. 89, 341–357. doi: 10.1002/ana.25962

Reinoso, G., Allen, J. C. Jr, Au, W. L., Seah, S. H., Tay, K. Y., and Tan, L. C. (2015). Clinical evolution of Parkinson's disease and prognostic factors affecting motor progression: 9-year follow-up study. *Eur. J. Neurol*. 22, 457–463. doi: 10.1111/ene.12476

Rossi-deVries, J., Pedoia, V., Samaan, M. A., Ferguson, A. R., Souza, R. B., and Majumdar, S. (2018). Using multidimensional topological data analysis to identify traits of hip OA. *J. Magn. Reson. Imaging*. 48, 1046–1058. doi: 10.1002/jmri.26029

Schrag, A., Siddiqui, U. F., Anastasiou, Z., Weintraub, D., and Schott, J. M. (2017). Clinical variables and biomarkers in prediction of cognitive impairment in patients with newly diagnosed Parkinson's disease: a cohort study. *Lancet Neurol*. 16, 66–75. doi: 10.1016/S1474-4422(16)30328-3

Selikhova, M., Williams, D. R., Kempster, P. A., Holton, J. L., Revesz, T., and Lees, A. J. (2009). A clinico-pathological study of subtypes in Parkinson's disease. *Brain* 132 (Pt. 11), 2947–2957. doi: 10.1093/brain/awp234

Severson, K. A., Chahine, L. M., Smolensky, L. A., Dhuliawala, M., Frasier, M., Ng, K., et al. (2021). Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. *Lancet Digit. Health*. 3, e555-e564. doi: 10.1016/S2589-7500(21)00101-1

Simuni, T., Long, J. D., Caspell-Garcia, C., Coffey, C. S., Lasch, S., Tanner, C. M., et al. (2016). Predictors of time to initiation of symptomatic therapy in early Parkinson's disease. *Ann. Clin. Transl. Neurol*. 3, 482–494. doi: 10.1002/acn3.317

Singh, G., Memoli, F., and Carlsson, G. E. (2007). "Topological methods for the analysis of high dimensional data sets and 3d object recognition," in *Eurographics Symposium on Point-Based Graphics*, 91–100.

van Veen, H. J., Saul, N., Eargle, D., and Mangham, S. W. (2019a). Kepler mapper: a flexible python implementation of the mapper algorithm. *J. Open Sourc. Softw*. 4, 1315. doi: 10.21105/joss.01315

van Veen, H. J., Saul, N., Eargle, D., and Mangham, S. W. (2019b). *Kepler Mapper: A Flexible Python Implementation of the Mapper Algorithm (Version 1.4.1)*. Zenodo. doi: 10.5281/zenodo.4077395

Velseboer, D. C., Broeders, M., Post, B., van Geloven, N., Speelman, J. D., Schmand, B., et al. (2013). Prognostic factors of motor impairment, disability, and quality of life in newly diagnosed PD. *Neurology*. 80, 627–633. doi: 10.1212/WNL.0b013e318281cc99

Vu, T. C., Nutt, J. G., and Holford, N. H. (2012). Progression of motor and nonmotor features of Parkinson's disease and their response to treatment. *Br. J. Clin. Pharmacol*. 74, 267–283. doi: 10.1111/j.1365-2125.2012.04192.x