Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Identifying Parkinson's disease subtypes with motor and non-motor symptoms via model-based multi-partition clustering

Fernando Rodriguez-Sanchez (fernando.rodriguezs@upm.es)
Universidad Politenica de Madrid

Carmen Rodriguez-Blazquez
National Center of Epidemiology, Carlos III Institute of Health

Concha Bielza
Universidad Politenica de Madrid

Pedro Larrañaga
Universidad Politenica de Madrid

Daniel Weintraub
University of Pennsylvania

Pablo Martinez-Martin
Carlos III Institute of Health

Alexandra Rizos
King's College London, King's College Hospital

Anette Schrag
University College London

K. Ray Chaudhuri
King's College London, King's College Hospital

Research Article

Keywords: multi-center, sex, age, age of onset, disease duration, Hoehn & Yahr stage, clusters, latent class analysis, Gaussian mixture mode, MDS-UPDRSI

Posted Date: August 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-828016/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Scientific Reports on December 1st, 2021. See the published version at https://doi.org/10.1038/s41598-021-03118-w.

Identifying Parkinson's disease subtypes with motor and non-motor symptoms via model-based multi-partition clustering

Fernando Rodriguez-Sanchez[1,*], Carmen Rodriguez-Blazquez[2,3], Concha Bielza[1], Pedro Larrañaga[1], Daniel Weintraub[4], Pablo Martinez-Martin[3], Alexandra Rizos[5], Anette Schrag[6], and K. Ray Chaudhuri[5]

[1]Computational Intelligence Group, Universidad Politécnica de Madrid, Madrid, Spain.
[2]National Center of Epidemiology, Carlos III Institute of Health, Madrid, Spain.
[3]Center for Networked Biomedical Research in Neurodegenerative Diseases (CIBERNED), Carlos III Institute of Health. Madrid, Spain.
[4]Departments of Psychiatry and Neurology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, United State.
[5]King's College London, Department of Neurosciences, Institute of Psychiatry, Psychology & Neuroscience and Parkinson's Foundation Centre of Excellence, King's College Hospital, London, UK.
[6]Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London, UK.
[*]email: fernando.rodriguezs@upm.es

ABSTRACT

Identification of Parkinson's disease subtypes may help understand underlying disease mechanisms and provide personalized management. Although clustering methods have been previously used for subtyping, they have reported generic subtypes of limited relevance in real life practice because patients do not always fit into a single category. The aim of this study was to identify new subtypes assuming that patients could be grouped differently according to certain sets of related symptoms. To this purpose, a novel model-based multi-partition clustering method was applied on data from an international, multi-center, cross-sectional study of 402 Parkinson's disease patients. Both motor and non-motor symptoms were considered. As a result, eight sets of related symptoms were identified. Each of them provided a different way to group patients: impulse control issues, overall non-motor symptoms, presence of dyskinesias and pyschosis, fatigue, axial symptoms and motor fluctuations, autonomic dysfunction, depression, and excessive sweating. Each of these groups could be seen as a subtype of the disease. Significant differences between subtypes ($p < 0.01$) were found in sex, age, age of onset, disease duration, and Hoehn & Yahr stage. Independent confirmation of these results could have implications for the clinical management of Parkinson's disease patients.

Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disease that is clinically characterized by a broad spectrum of motor and non-motor manifestations[1]. There is, however, considerable clinical phenotypic and natural history related variability between PD patients, which may indicate the existence of disease subtypes. Identification of PD subtypes may help understand the underlying disease mechanisms, since homogeneous groups of patients may be more likely to share pathological and genetic features. In addition, identification of PD subtypes may ultimately lead to more precise treatment strategies (i.e., precision medicine)[2].

Data-driven techniques such as clustering may be suitable for establishing PD subtypes. In clustering, patients are assigned to several groups (i.e., clusters) so that patients belonging to the same group share similarities. Each of these groups is usually then considered a subtype of the disease. Previous clustering studies have already identified PD subtypes with motor and non-motor symptoms[3–8]. However, to the best of our knowledge, all of them have used single-partition clustering methods such as k-means[9], latent class analysis[10], Gaussian mixture model[11], agglomerative hierarchical clustering[12]. Single-partition clustering algorithms assume the existence of a single true clustering in a dataset. As a result, each patient is assigned to a single subtype that is defined by all the considered symptoms.

Recently, several issues have been raised about data-driven PD subtypes, such as the low number in the samples, their lack of internal homogeneity, and their difficulty to reproduce meaningful data in real life and external validity[13,14]. We believe that these issues may be a consequence of using single-partition clustering methods. The assumption that each patient should be assigned to a single generic subtype does not hold for PD, which is usually multifaceted and can be meaningfully partitioned in multiple ways[15,16]. For this reason, we advocate for model-based multi-partition clustering[17–20], which extends model-based clustering[11] by producing mixture models with multiple categorical latent variables. The idea is to use statistical principles to find sets of related symptoms where patients are divided into a number of distinct groups. Each set of symptoms defines a different clustering of patients. As a result, each patient is assigned to one subtype for each clustering. The analysis of these subtypes and their associations may provide more accurate insights about the considered symptoms, as well as their relationship with socio-demographic and clinical information of the patients.

Based on the above, the objectives of our study were: (i) to identify PD subtypes using model-based multi-partition clustering, and (ii) to analyze the associations between the resulting subtypes. To this end, we developed a novel model-based multi-partition clustering algorithm, and applied this method on data from a large, multi-center, international, and well-characterized cohort of patients.

Methods

Data

The analysis was carried out on data gathered from the first validation study of the Movement Disorder Society Non-Motor Rating Scale (MDS-NMS), an international, multi-center, cross-sectional study that included PD English-speaking patients from England and the United States[21]. The study was approved by the institutional review boards or ethics committees of the participating centers. All patients gave their written informed consent to participate in the study. Institutional review boards or ethics committees that approved the study: (1) National Research Ethics Service Committee East Midlands - Northampton, England; (2) Institutional Review Board at the Perelman School of Medicine at the University of Pennsylvania, United States. In addition, the study was conducted according to good clinical practice and all research was performed in accordance with relevant guidelines and regulations. Data are publicly available in our GitHub repository[22]. For all patients, socio-demographic information and basic clinical variables (i.e., sex, age, age of onset, and disease duration) were recorded and the following assessments were applied:

1. The Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS)[23], which is composed of 65 items divided across 4 parts, namely, Part I: Non-motor Experiences of Daily Living (13 items); Part II: Motor Experiences of Daily Living (13 items); Part III: Motor Examination (33 items); and Part IV: Motor Complications (6 items). Each item has 5 options of response, running from 0 (normal) to 4 (maximum intensity). The total score of each part is obtained by summing the respective item scores.

2. The MDS-NMS[21], which is composed of 52 items grouped into 13 domains: depression, anxiety, apathy, psychosis, impulse control and related disorders (ICRDs), cognition, orthostatic hypotension, urinary, sexual, gastrointestinal, sleep and wakefulness, pain, and other. Each item is scored for both frequency and severity, where both scores have 5 options of response, ranging from 0 (normal) to 4 (maximum intensity). Each item score is generated by multiplying frequency and severity. The score of each domain is obtained by summing the respective item scores. The MDS-NMS also includes a subscale for non-motor fluctuations, composed of 8 items, which was not considered in this study.

3. The Hoehn & Yahr (H&Y) staging system[24], which ranges from 1 to 5.

Motor items from the MDS-UPDRS were classed as 5 motor cardinal signs: tremor, rigidity, bradykinesia, dyskinesias and motor fluctuations; and 2 motor subtypes: axial symptoms and postural instability gait difficulty (PIGD)[25]. This resulted in 7 motor variables. The specific MDS-UPDRS items that constitute each motor variable are provided in Section 1 of the Supplementary Information. Additionally, items from the MDS-NMS were grouped into their respective domains, with the exception of the items from the "other" domain (unintentional weight loss, decrease in sense of smell, physical fatigue, mental fatigue, and excessive sweating). These items were individually considered due to their individual and unique status. This resulted in 17 non-motor variables. Finally, with the objective of improving the interpretability of the results, both motor and non-motor variables were normalized to the [0, 1] range using min-max scaling.

A total of 402 patients were considered for this study. Average onset age was 59 ± 11 (s.d.) years, 62% were male and average PD duration was 8 ± 6 (s.d.) years. 13% of the patients were in H&Y stage 1; 54% in H&Y stage 2; 28% in H&Y stage 3; and 5% in H&Y stage 4. No patients in this study were in H&Y stage 5. With respect to missing information, 64 values ($< 1\%$ of the total) were missing, mostly in the Sexual domain of the MDS-NMS. As our multi-partition clustering method was able to work with missing information, no patients were excluded from the analysis. For more information about the data, see Table 1.

	N[1]	Mean	(s.d.)	Median	Min[2]	Max[2]
Age	402	67.42	(9.96)	68	35	93
Age of onset	402	59.23	(10.67)	59	26	93
Disease duration	402	8.19	(5.93)	7	0	35
H&Y	402	2.25	(0.74)	2	1	4
MDS-UPDRS						
Tremor	391	0.13	(0.12)	0.00	0.11	0.57
Rigidity	398	0.19	(0.16)	0.00	0.15	1.00
Dyskinesias	402	0.07	(0.15)	0.00	0.00	1.00
Motor fluctuations	401	0.16	(0.19)	0.00	0.06	0.81
Bradykinesia	394	0.29	(0.17)	0.00	0.25	0.89
Axial symptoms	402	0.23	(0.15)	0.00	0.21	0.86
PIGD	393	0.22	(0.19)	0.00	0.15	0.85
MDS-NMS						
Depression	401	0.07	(0.13)	0.01	0.00	0.90
Anxiety	402	0.09	(0.13)	0.03	0.00	0.84
Apathy	402	0.08	(0.15)	0.00	0.00	0.75
Psychosis	402	0.03	(0.06)	0.00	0.00	0.56
ICRDs	401	0.02	(0.05)	0.00	0.00	0.39
Cognition	402	0.10	(0.12)	0.05	0.00	0.69
Orthostatic hypotension	402	0.07	(0.13)	0.00	0.00	0.75
Urinary	402	0.16	(0.19)	0.08	0.00	1.00
Sexual	375	0.14	(0.25)	0.00	0.00	1.00
Gastrointestinal	401	0.10	(0.12)	0.06	0.00	0.73
Sleep and wakefulness	401	0.12	(0.12)	0.08	0.00	0.79
Pain	402	0.13	(0.15)	0.08	0.00	0.83
Unintentional weight loss	402	0.06	(0.18)	0.00	0.00	1.00
Decrease in sense of smell	402	0.39	(0.40)	0.25	0.00	1.00
Physical fatigue	402	0.21	(0.27)	0.06	0.00	1.00
Mental fatigue	402	0.09	(0.19)	0.00	0.00	1.00
Excessive sweating	402	0.07	(0.18)	0.00	0.00	1.00

Table 1. Descriptive statistics of the data. Numbers between parentheses correspond to standard deviations (s.d.). [1]Sample size without missing values. [2]Min and max recorded values.

Model-based multi-partition clustering method

A novel model-based multi-partition clustering method was developed to identify groups of individuals with specific patterns in the motor and non-motor domains. The proposed method learned a conditional linear Gaussian Bayesian network (BN)[26] with multiple categorical latent variables. Each latent variable provided a unique way to partition PD patients according to a unique set of symptom variables. Each group of patients was considered a PD subtype with respect to the partition variables.

Two components can be distinguished in every BN: (i) a directed acyclic graph that encodes conditional independences among triplets of variables in the BN; and (ii) a set of parameters that describe the conditional probability distributions of each variable given its parents in the graph. Together, both of these elements define a unique joint probability distribution. BNs are useful in multi-partition clustering for several reasons. First, their graphical structure allows for an easier interpretation, showing which variables define each partition, and how partitions relate to each other. Second, their conditional independences result in more compact models that are easier to learn from data. Finally, BNs allow probabilistic inference, which is useful for making predictions, diagnoses and explanations.

Our proposal iteratively explores the space of conditional linear Gaussian BNs using five latent operators and a variational Bayesian[27] version of the structural expectation-maximization[28] algorithm. Latent operators are tasked with introducing latent variables, removing latent variables, and changing the cardinality (i.e., number of subtypes) of latent variables. Each application of these operators produces a candidate model whose structure is refined using the variational Bayesian structural expectation-maximization algorithm. Once all the candidate models have been evaluated, the highest scoring model is selected. This process is iteratively repeated until the model score ceases to increase. Given its greedy nature, we refer to this method as greedy latent structure learner. It is formally defined in Sections 2 and 3 of the Supplementary Information, and its implementation in Java 8

is publicly available in the project's Github repository[22].

Analysis of multi-partition PD subtypes

By using a conditional linear Gaussian BN, each subtype in a partition was defined by a linear Gaussian distribution whose dimensions corresponded to the partition symptoms. The symbol μ was used to denote the mean of this subtype for a specific symptom and the symbol σ was used to denote the s.d. In addition, to improve the readability of these subtypes, we devised a simple scale that considered the quartiles of the normalized [0, 1] range to refer to the mean symptom severity: (i) slight [0.01, 0.25]; (ii) mild [0.26, 0.50]; (iii) moderate [0.51, 0.75]; and (iv) severe [0.76, 1]. Note that this scale differs from the MDS-UPDRS and MDS-NMS ratings.

To explore the relationship between socio-demographic information, basic clinical variables, H&Y stage and the identified subtypes, hypothesis tests were performed. Each pair of subtypes in a partition were compared. For continuous variables such as age, age of PD onset, and PD duration, an ANOVA test or a Mann-Whitney U-test (both implemented in the Python library SciPy version 1.5.2) was used. When three or more groups were present in a clustering, an ANOVA test or a Kruskal-Wallis test (both implemented in SciPy) was performed instead, followed by a post-hoc analysis using Tukey's range test (implemented in the Python library Statsmodels version 0.11.1). For categorical variables such as the sex of the patient and discrete variables such as the H&Y stage, χ^2 tests (implemented in SciPy) were performed. Statistical significance was defined as p-value $p < 0.01$.

To analyze the associations between the identified subtypes, probabilistic inference was employed. For example, consider a hypothetical multi-partition clustering model with two partitions, A and B, which are connected by an arc in the model. Partition A defines two PD subtypes $\{A1, A2\}$ according to a set of symptoms. Partition B defines three PD subtypes $\{B1, B2, B3\}$ according to a different set of symptoms. We are interested in estimating the difference between the probability distributions $P(B)$ and $P(B|A = A1)$, but also the difference between $P(B)$ and $P(B|A = A2)$. That is, how being assigned a subtype in A affects the probability distribution of B. The inverse probabilistic queries are also relevant (i.e., how being assigned a subtype in B affects the probability distribution of A). Since each subtype in a partition is characterized by a set of symptoms with a certain severity, we are incidentally studying the relationships between their respective symptoms (i.e., how an increase or decrease of the severity of certain symptoms affect the probability of suffering the other symptoms with more or less severity) when we analyze the relationships between subtypes of different partitions. In this study, probabilistic queries were carried out using Monte Carlo sampling in the tool for BN analysis GeNIe (version 3.0).

Results

Multi-partition PD subtypes

The BN structure that resulted from applying our multi-partition clustering algorithm on the 7 motor and 17 non-motor variables is portrayed in Fig. 1. It consisted of 9 (alphabetically-named) latent variables. Each latent variable defined a unique partition according its descendant symptom variables in the graph. For example, in partition A, patients were divided into two subtypes according to the severity of their ICRDs and PIGD. There was, however, one latent variable that differed from the rest by not being directly related to any symptom variable. Instead, it acted as an auxiliary latent variable that connected partitions G (weight loss - depression) and H (excessive sweating - anxiety). This variable was I, and to simplify the analysis, its relevant information was condensed in those of G and H. As a result, 8 partitions were discovered, each with a different number of subtypes. The sex, age, age of onset, disease duration, and H&Y stage of each subtype is provided in Table 2. Significant differences between subtypes are included. Results (i.e., p-values) of the statistical tests that were performed are provided in Section 4 of the Supplementary Information.

Partition A (ICRDs - PIGD)

- Subtype $A1$ (84%) was characterized by patients that did not show problems to control their impulses ($\mu = 0.00$, $\sigma = 0.00$), but did show slight PIGD ($\mu = 0.22$, $\sigma = 0.19$).
- Subtype $A2$ (16%) was characterized by patients that showed slight problems to control their impulses ($\mu = 0.09$, $\sigma = 0.08$), and also presented slight PIGD ($\mu = 0.20$, $\sigma = 0.16$).

Partition B (apathy - cognitive - pain - gastrointestinal - sleep - urinary)

- Subtype $B1$ (49%) was formed of patients that showed no apathy ($\mu = 0.00$, $\sigma = 0.00$), slight cognitive changes ($\mu = 0.03$, $\sigma = 0.04$), slight pain ($\mu = 0.06$, $\sigma = 0.07$), slight gastrointestinal problems ($\mu = 0.04$, $\sigma = 0.05$), slight sleep disorders ($\mu = 0.06$, $\sigma = 0.07$), and slight urinary issues ($\mu = 0.06$, $\sigma = 0.09$).
- Subtype $B2$ (51%) was formed of patients that showed slight apathy ($\mu = 0.16$, $\sigma = 0.18$), slight cognitive changes ($\mu = 0.17$, $\sigma = 0.14$), slight pain ($\mu = 0.19$, $\sigma = 0.17$), slight gastrointestinal problems ($\mu = 0.15$, $\sigma = 0.14$), slight sleep disorders ($\mu = 0.17$, $\sigma = 0.13$), and slight urinary issues ($\mu = 0.25$, $\sigma = 0.22$).

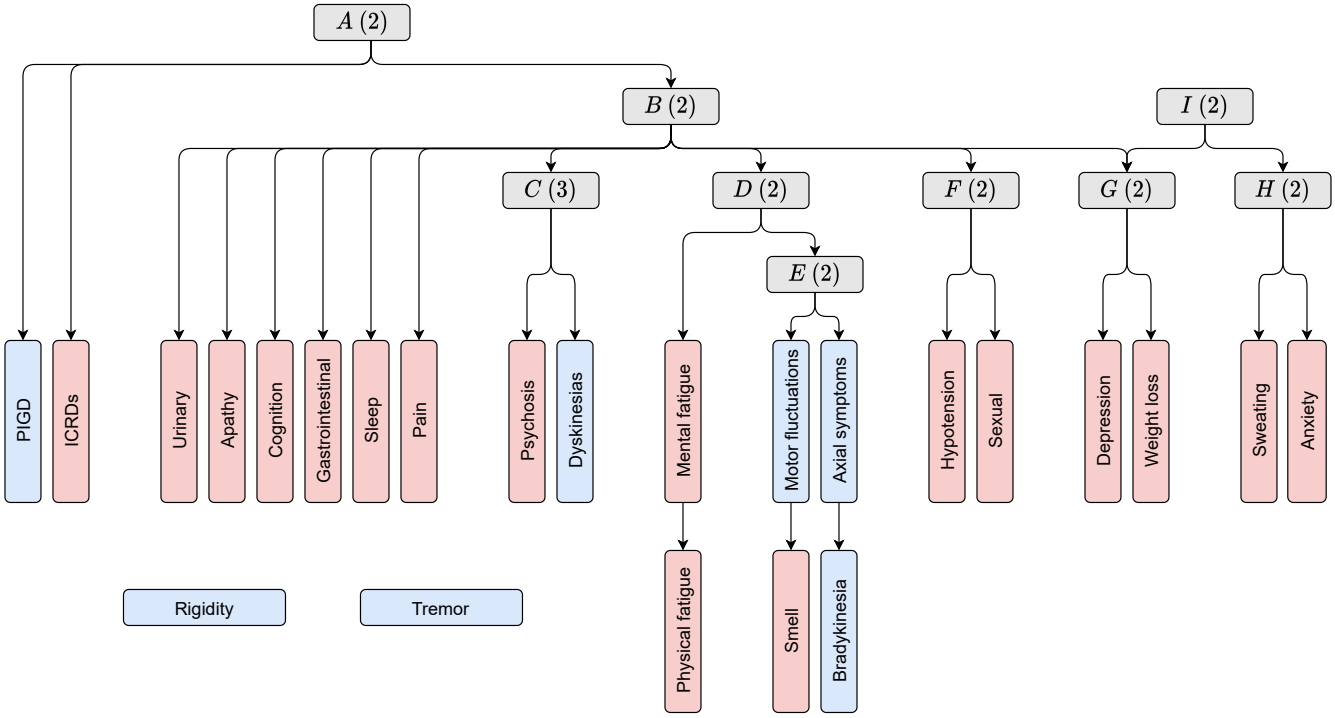

Figure 1. BN structure of the multi-partition clustering model. Blue nodes represent motor variables and red nodes represent non-motor variables. Grey nodes represent categorical latent variables (i.e., partitions), where the number in parentheses corresponds to the variable's cardinality (i.e., the number of subtypes). Graph arcs represent conditional dependences.

	Subtype	Sex (% male)	Age	Age of onset	Duration	H&Y
Partition A	$A1$	62.39	68.02 (9.85)^{a}	60.00 (10.82)^{a}	8.02 (5.95)	2.27 (0.76)
	$A2$	61.19	64.45 (10.02)^{a}	55.37 (9.06)^{a}	9.07 (5.79)	2.16 (0.62)
Partition B	$B1$	60.80	66.07 (10.18)^{b}	59.18 (10.79)	6.89 (5.15)^{b}	2.05 (0.71)^{b}
	$B2$	63.55	68.75 (9.58)^{b}	59.28 (10.59)	9.48 (6.36)^{b}	2.45 (0.72)^{b}
Partition C	$C1$	60.74	67.42 (10.19)	60.72 (10.70)^{c_1,c_2}	6.70 (5.35)^{c_1,c_2}	2.10 (0.73)^{c_1,c_2}
	$C2$	64.29	67.10 (9.83)	56.98 (11.00)^{c_1}	10.12 (6.19)^{c_1}	2.46 (0.70)^{c_1}
	$C3$	64.58	68.19 (9.22)	56.92 (8.37)^{c_2}	11.27 (5.73)^{c_2}	2.54 (0.71)^{c_2}
Partition D	$D1$	61.76	68.17 (10.21)	60.48 (10.92)^{d}	7.68 (5.65)^{d}	2.17 (0.74)
	$D2$	63.08	65.87 (9.27)	56.60 (9.66)^{d}	9.27 (6.36)^{d}	2.42 (0.72)
Partition E	$E1$	63.32	68.61 (9.75)^{e}	62.25 (10.26)^{e}	6.36 (5.18)^{e}	2.13 (0.74)^{e}
	$E2$	61.08	66.26 (10.05)^{e}	56.27 (10.26)^{e}	10.00 (6.08)^{e}	2.37 (0.72)^{e}
Partition F	$F1$	54.67^{f}	66.64 (10.49)	59.29 (11.04)	7.35 (5.29)^{f}	2.11 (0.75)^{f}
	$F2$	70.74^{f}	68.31 (9.26)	59.15 (10.27)	9.16 (6.46)^{f}	2.41 (0.69)^{f}
Partition G	$G1$	63.09	67.37 (10.16)	59.31 (10.86)	8.05 (5.93)	2.20 (0.74)
	$G2$	58.82	67.74 (9.24)	58.91 (10.02)	8.73 (5.93)	2.46 (0.72)
Partition H	$H1$	63.55	68.25 (9.67)^{h}	60.38 (10.31)^{h}	7.87 (5.99)^{h}	2.23 (0.76)
	$H2$	58.25	65.02 (10.44)^{h}	55.87 (11.06)^{h}	9.15 (5.67)^{h}	2.31 (0.69)

Table 2. Sex, age, age of onset, disease duration and H&Y stage of each PD subtype. Numbers between parentheses correspond to standard deviations. ^{a}Significant differences between $A1$ and $A2$. ^{b}Significant differences between $B1$ and $B2$. ^{c_1}Significant differences between $C1$ and $C2$. ^{c_2}Significant differences between $C1$ and $C3$. ^{d}Significant differences between $D1$ and $D2$. ^{e}Significant differences between $E1$ and $E2$. ^{f}Significant differences between $F1$ and $F2$. ^{h}Significant differences between $H1$ and $H2$. Statistical significance was defined as p-value $p < 0.01$.

Partition C (dyskinesias - psychosis)

- Subtype $C1$ (59%) was composed of patients that showed no dyskinesias ($\mu = 0.00$, $\sigma = 0.01$) or psychosis ($\mu = 0.00$, $\sigma = 0.00$).
- Subtype $C2$ (27%) was composed of patients that showed slight dyskinesias ($\mu = 0.18$, $\sigma = 0.18$) and slight psychosis ($\mu = 0.02$, $\sigma = 0.03$).
- Subtype $C3$ (14%) was composed of patients that showed slight dyskinesias ($\mu = 0.15$, $\sigma = 0.23$) and slight psychosis ($\mu = 0.14$, $\sigma = 0.11$).

Partition D (mental fatigue - physical fatigue)

- Subtype $D1$ (66%) consisted of patients that showed no mental fatigue ($\mu = 0.00$, $\sigma = 0.01$) and slight physical fatigue ($\mu = 0.14$, $\sigma = 0.06$).
- Subtype $D2$ (34%) consisted of patients that showed mild mental fatigue ($\mu = 0.28$, $\sigma = 0.23$) and mild physical fatigue ($\mu = 0.35$, $\sigma = 0.06$).

Partition E (axial symptoms – bradykinesia – loss of smell – motor fluctuations)

- Subtype $E1$ (47%) was constituted by patients that showed slight axial symptoms ($\mu = 0.19$, $\sigma = 0.15$), slight bradykinesia ($\mu = 0.21$, $\sigma = 0.02$), mild loss of smell ($\mu = 0.29$, $\sigma = 0.15$), but no motor fluctuations ($\mu = 0.00$, $\sigma = 0.02$).
- Subtype $E2$ (53%) was constituted by patients that showed mild axial symptoms ($\mu = 0.26$, $\sigma = 0.15$), mild bradykinesia ($\mu = 0.30$, $\sigma = 0.02$), moderate loss of smell ($\mu = 0.51$, $\sigma = 0.15$), and mild motor fluctuations ($\mu = 0.30$, $\sigma = 0.16$).

Partition F (orthostatic hypotension - sexual problems)

- Subtype $F1$ (55%) was composed of patients that showed no orthostatic hypotension ($\mu = 0.00$, $\sigma = 0.02$) and slight sexual problems ($\mu = 0.01$, $\sigma = 0.03$).
- Subtype $F2$ (45%) was composed of patients that showed slight orthostatic hypotension ($\mu = 0.15$, $\sigma = 0.17$) and mild sexual problems ($\mu = 0.29$, $\sigma = 0.30$).

Partition G (weight loss - depression)

- Subtype $G1$ (80%) was characterized by patients that showed no weight loss ($\mu = 0.00$, $\sigma = 0.01$) and slight depression ($\mu = 0.03$, $\sigma = 0.04$).
- Subtype $G2$ (20%) was characterized by patients that showed mild weight loss ($\mu = 0.26$, $\sigma = 0.32$) and slight depression ($\mu = 0.24$, $\sigma = 0.20$).

Partition H (excessive sweating - anxiety)

- Subtype $H1$ (74%) consisted of patients that showed no degree of excessive sweating ($\mu = 0.00$, $\sigma = 0.01$) and slight anxiety ($\mu = 0.06$, $\sigma = 0.07$).
- Subtype $H2$ (26%) consisted of patients that showed mild degree of excessive sweating ($\mu = 0.27$, $\sigma = 0.28$) and slight anxiety ($\mu = 0.19$, $\sigma = 0.19$).

A total of 29 probabilistic queries were performed to analyze the connections between the identified subtypes. They are provided in Section 4 of the Supplementary Information.

Comparison with other model-based clustering methods.

We compared our model-based multi-partition clustering method with two model-based single-partition clustering methods (i.e., the latent class model[10], the Gaussian mixture model[11], and the unsupervised k-dependence Bayesian classifier[29]), and two model-based multi-partition clustering methods (i.e., the Gaussian expansion simplification until termination algorithm[17], and the multi-partition mixture model[18]). We evaluated the quality of the results from both a data fitting and a clustering perspective.

In this comparative analysis, we observed that multi-partition clustering methods were able to obtain multiple partitions from data, which resulted in a higher number of subtypes than single-partition clustering methods. These subtypes were not only more specific, but also more faithful to the data (i.e., higher model selection scores). From the considered methods, our approach returned the highest scoring model. The Gaussian expansion simplification until termination algorithm also obtained a high model selection score. However, its model suffered from overfitting and was difficult to interpret (it identified 18 partitions with 55 subtypes). For more information about the model selection process, see Section 4 of the Supplementary Information.

Discussion

Clinical interpretation of PD subtypes and their associations

Partitions were underpinned by a reasonable spread of contributory PD symptoms, thus bridging a statistical and clinical divide. Tremor and rigidity were the exceptions, appearing to be independent of the rest of variables in the model (see Fig. 1). Weak correlation between rigidity, tremor, and non-motor symptoms is not uncommon and has also been observed in a recent study that considered a similar population[30].

In partition *A*, patients were divided into two subtypes according to the severity of their ICRDs and PIGD. Although the mean PIGD of the subtypes did not differ by much, subtype *A*2 was characterized by a higher severity of ICRDs, a younger age and a younger age of onset. A relationship between young age, early PD onset and more severe ICRDs has been previously observed[31]. Both socio-demographic aspects are known risk factors for ICRDs along with motor complications, a pre-PD history of ICRDs, and a dopamine agonist treatment[32].

Apathy, cognition, pain, gastrointestinal, sleep, and urinary symptoms were associated in partition *B*. Two subtypes were identified, where patients characterized as subtype *B*2 presented a higher severity of these symptoms. This subtype is consistent with the Parkinson's apathy subtype[33,34], which has been described to be formed by older patients that showed cognitive impairment, sleep disorders, and relatively severe motor symptoms. The relationship between sleep disorders and urinary problems may indicate the presence of nocturia[35]. In addition, a recent study has also identified a relationship between constipation and cognitive dysfunction in two independent cohorts of patients[36].

Partition *C* distinguished three subtypes that differed according to the severity of dyskinesias and psychosis. Subtypes *C*2 and *C*3 consisted of patients with a longer duration of the disease and a younger age of onset. These subtypes coincided with the observation that dyskinesias and psychosis are usually present in late stages of PD and may be associated with higher dopaminergic medication doses[37,38].

Fatigue is considered a common and complex non-motor symptom of PD, prevalent from the prodromal to the palliative stage. It is usually present from early stages of the disease and may often persist or even worsen over time[39]. While fatigue is usually regarded as an independent symptom, it has been moderately associated with apathy, sleep disorders, depression, and motor problems[40,41]. Our model was able to capture this duality by identifying a specific partition for fatigue symptoms, and connecting it with partitions *B* (apathy, sleep and depression) and *E* (motor problems). In addition, patients that suffered from more severe fatigue showed a longer duration of the disease and a younger age of PD onset.

Bradykinesia, axial symptoms, and motor fluctuations were associated in partition *E* with a decrease in sense of smell (i.e., hyposmia). Patients were divided into two subtypes according to *E*. While both subtypes presented motor issues, *E*2 was characterized by a higher severity of motor symptoms, hyposmia, and the presence of motor fluctuations. Anosmia/hyposmia is considered a preclinical marker of PD with relatively static severity. However, while it has not been associated to any particular PD phenotype[42,43], a recent study has noted that normosmic PD patients usually display better motor function than hyposmic PD patients[44].

Partition *F* identified two subtypes based on orthostatic hypotension and sexual problems. While the rest of partitions were independent of the sex of the patient, 71% of patients in *F*2 were male, showing significant differences in the sex of the patients belonging to *F*1 and *F*2. We also observed significant differences in the H&Y stage and PD duration of these patients, reflecting the later occurrence of the autonomic features of orthostatic hypotension and sexual dysfunction[45–47].

Weight loss and depression were associated in partition *G*. Two subtypes were identified, where *G*2 was characterized by patients with mild weight loss and depression. Loss of appetite due to depression is a known weight loss factor[48]. There were no significant differences in sex, age, age of onset or H&Y stage of the patients belonging to *G*1 and *G*2.

Regarding clustering *H*, anxiety was associated with excessive sweating. Anxiety was present in both the *H*1 and *H*2 subtypes, but patients in *H*2 showed a higher severity of this symptom along with a mild degree of excessive sweating. This result is consistent with a recent study[49] in which anxiety and depression were more prominent in PD patients with hyperhidrosis.

It is important to note that none of the discovered subtypes were fully independent of each other. Belonging to a specific subtype in a partition influenced the subtype probabilities in the rest of partitions. By using probabilistic inference, we were able to study the effect of these associations on their respective symptoms. Some interesting patterns that we observed included: (i) patients with ICRDs (*A*2 subtype) had a 0.75 probability of presenting the symptoms of *B*2. This result is consistent with a recent study that has challenged the traditional concept of apathy and ICRDs as opposite symptoms[50]; (ii) patients that suffered psychosis (*C*3 subtype) had an 0.88 probability of suffering the symptoms of *B*2. The presence of visual hallucinations has been linked to sleep deprivation, cognitive impairment and depression[51,52]; and (iii) patients with mild mental and physical fatigue (*D*2 subtype) had a 0.79 probability of suffering the symptoms of *B*2 and a 0.73 probability of suffering the symptoms of *E*2. As previously discussed, fatigue has been related to the presence of apathetic symptoms, sleep disturbances, and higher H&Y stages[40,41].

The majority of partitions were directly or indirectly influenced by *B*, which acted as a pivotal latent variable in the

multi-partition clustering model. This aligns with the current observation that sleep disorders, depression, constipation, and other non-motor symptoms appear across the spectrum of patients with PD[53].

Limitations

This study has some limitations. Concerning the population of the study, patients were not specifically selected for this analysis, but rather for the validation of the MDS-NMS. Nonetheless, the large sample size and the high quality of the collected data will allow these results to be contrasted and compared with the results of future studies. The sample was comparatively younger than the average population of patients with PD. It is therefore possible that the results differ in those with an older age where higher rates of comorbidities exist. In addition, we did not report a control group, although our intention was not to describe the symptoms as discriminant from normal subjects. Finally, we did not consider PD biomarkers, which could provide more information about the identified subtypes.

Conclusion

Dividing PD patients into groups with common symptoms may help understand their underlying pathological processes. In this study, we used model-based multi-partition clustering to categorize patients according to 8 different sets of motor and/or non-motor symptoms. By using probabilistic inference, we were able to explore the associations between these subtypes and extract useful patterns. Independent confirmation of these results could allow for more precise PD treatments. In the future, it would be interesting to research how the evolution of PD throughout the years would affect these subtypes, and to which extent they could be markers of PD progression.

References

1. Greenland, J. C., Williams-Gray, C. H. & Barker, R. A. The clinical heterogeneity of Parkinson's disease and its therapeutic implications. *Eur. J. Neurosci.* **49**(3), 328-338 (2019).
2. Marras, C., Chaudhuri, K. R., Titova, N. & Mestre, T. A. Therapy of Parkinson's disease subtypes. *Neurotherapeutics* **17**, 1366-1377 (2020).
3. Van Rooden, S. M. *et al*. Clinical subtypes of Parkinson's disease. *Mov. Disord.* **26**(1), 51-58 (2011).
4. Ma, L. Y., Chan, P., Gu, Z. Q., Li, F. F. & Feng, T. Heterogeneity among patients with Parkinson's disease: Cluster analysis and genetic association. *J. Neurol. Sci.* **351**(12), 41-45 (2015).
5. Erro, R. *et al.* Clinical clusters and dopaminergic dysfunction in de-novo Parkinson disease. *Parkinsonism Relat. Disord.* **28**, 137-140 (2016).
6. Fereshtehnejad, S. M. & Postuma, R. B. Subtypes of Parkinson's disease: What do they tell us about disease progression?. *Curr. Neurol. Neurosci. Rep.* **17**(4), 34 (2017).
7. Mu, J. *et al.* Parkinson's disease subtypes identified from cluster analysis of motor and non-motor symptoms. *Front. Aging Neurosci.* **9**, 301 (2017).
8. Martinez-Martin, P. *et al.* Factor analysis and clustering of the movement disorder society non-motor rating scale. *Mov. Disord.* **35**(6), 969-975 (2020).
9. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (BSMSP)* 281-297 (1967).
10. Lazarsfeld, P. F. & Henry, N. W. *Latent Structure Analysis*. (Hougton & Mifflin, 1968).
11. McLachlan, G. J., Lee, S. X. & Rathnayake, S. I. Finite mixture models. *Annu. Rev. Stat. Appl.* **6**(1), 355-378 (2019).
12. Gordon, A. D. A review of hierarchical classification. *J. R. Stat. Soc. (Series A Stat. Soc.)* **150**(2), 119-137 (1987).
13. Qian, E. & Huang, Y. Subtyping of Parkinson's disease - where are we up to?. *Aging Dis.* **10**(5), 1130 (2019).
14. Mestre, T. A. *et al.* Parkinson's disease subtypes: Critical appraisal and recommendations. *J. Parkinsons Dis.* **11**(2), 395-404 (2021).
15. Thenganatt, M. A. & Jankovic, J. Parkinson disease subtypes. *JAMA Neurol.* **71**(4), 499–504 (2014).
16. Marras, C. Subtypes of Parkinson's disease: State of the field and future directions. *Curr. Opin. Neurol.* **28**(4), 382-386 (2015).
17. Poon, L. K., Zhang, N. L., Liu, T. & Liu, A. H. Model-based clustering of high-dimensional data: Variable selection versus facet determination. *Int. J. Approx. Reason.* **54**(1), 196-215 (2013).

18. Galimberti, G., Manisi, A. & Soffritti, G. Modelling the role of variables in model-based cluster analysis. *Stat. Comput.* **28**(1), 145-169 (2018).

19. Li, X., Chen, Z., Poon, L. K. & Zhang, N. L. Learning latent superstructures in variational autoencoders for deep multidimensional clustering. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)* 1-15 (2019).

20. Falck, F. *et al.* Multi-facet clustering variational autoencoders. Preprint at https://arxiv.org/abs/2106.05241 (2021).

21. Chaudhuri, K. R. *et al.* The Movement Disorder Society Non-Motor Rating Scale (MDS-NMS): Initial validation study. *Mov. Disord.* **35**(1), 116-123 (2019).

22. Rodriguez-Sanchez *et al.* GitHub repository. https://github.com/ferjorosa/parkinson-subtypes (2021).

23. 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**(15), 2129-2170 (2008).

24. Hoehn, M. M. & Yahr, M. D. Parkinsonism: Onset, progression and mortality. *Neurol.* **17**, 427-442 (1967).

25. 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**(5), 668-670 (2013).

26. Lauritzen, S. L. & Wermuth, N. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Stat.* **17**, 31–57 (1989).

27. Attias, H. A variational Bayesian framework for graphical models. In *Proceedings of the 14th Conference on Neural Information Processing Systems (NIPS)* 209-215 (2000).

28. Friedman, N. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the 14th International Conference on Machine Learning (ICML)* 125-133 (1997).

29. Pham, D. T. & Ruz, G. A. Unsupervised training of Bayesian networks for data clustering. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **465**(2109), 2927-2948 (2009).

30. Berganzo, K. *et al.* Motor and non-motor symptoms of Parkinson's disease and their impact on quality of life and on different clinical subgroups. *Neurologia* **31**(9), 585-591 (2016).

31. Weintraub, D., David, A. S., Evans, A. H., Grant, J. E. & Stacy, M. Clinical spectrum of impulse control disorders in Parkinson's disease. *Mov. Disord.* **30**(2), 121-127 (2015).

32. Gatto, E. M. & Aldinio, V. Impulse control disorders in Parkinson's disease. A brief and comprehensive review. *Front. Neurol.* **10**, 351 (2019).

33. Dujardin, K. *et al.* Apathy in untreated early-stage Parkinson disease: Relationship with other non-motor symptoms. *Mov. Disord.* **29**(14), 1796-1801 (2014).

34. Sauerbier, A., Jenner, P., Todorova, A. & Chaudhuri K. R. Non-motor subtypes and Parkinson's disease. *Parkinsonism Relat. Disord.* **22**, 41-46 (2015).

35. Batla, A., Phé, V., Min, L. D. & Panicker, J. N. Nocturia in Parkinson's disease: Why does it occur and how to manage?. *Mov. Disord. Clin. Pract.* **3**(5), 443-451 (2016).

36. Leta, V. *et al.* Constipation is associated with development of cognitive Impairment in de novo Parkinson's disease: A longitudinal analysis of two international cohorts. *J. Parkinsons Dis.* **11**(3), 1209-1219 (2021).

37. Espay, A. J. *et al.* Levodopa-induced dyskinesia in Parkinson disease: Current and evolving concepts. *Ann. Neurol.* **84**(6), 797-811 (2018).

38. Dave, S., Weintraub, D., Aarsland, D. & Ffytche, D. H. Drug and disease effects in Parkinson's psychosis: revisiting the role of dopamine. *Mov. Disord. Clin. Pract.* **7**(1), 32-36 (2020).

39. Friedman, J. H. *et al.* Fatigue in Parkinson's disease: Report from a multidisciplinary symposium. *NPJ Parkinsons Dis.* **2**(1), 1-6 (2016).

40. Stocchi, F. *et al.* Prevalence of fatigue in Parkinson disease and its clinical correlates. *Neurol.* **83**(3), 215-220 (2014).

41. Siciliano, M. *et al.* Fatigue in Parkinson's disease: A systematic review and meta-analysis. *Mov. Disord.* **33**(11), 1712-1723 (2018).

42. Rossi, M. *et al.* Motor features in Parkinson's disease with normal olfactory function. *Mov. Disord.* **31**(9), 1414-1417 (2016).

43. Sui, X. *et al.* Hyposmia as a predictive marker of Parkinson's disease: A systematic review and meta-analysis. *Biomed. Res. Int.* **19**, 1-9 (2019).

44. Lee, D. H. *et al.* Is normosmic Parkinson disease a unique clinical phenotype?. *Neurol.* **85**(15), 1270-1275 (2015).

45. Özcan, T. *et al.* The association between symptoms of sexual dysfunction and age at onset in Parkinson's disease. *Clin. Auton. Res.* **26**(3), 205-209 (2016).

46. Palma, J. A. & Kaufmann H. Treatment of autonomic dysfunction in Parkinson disease and other synuclinopathies. *Mov. Disord.* **33**(3), 372-390 (2018).

47. Hiorth, Y. H., Pedersen, K. F., Dalen, I., Tysnes, O. B. & Alves G. Orthostatic hypotension in Parkinson disease: a 7-year prospective population-based study. *Neurol.* **93**(16), 1526-1534 (2019).

48. Ma, K. *et al.* Weight loss and malnutrition in patients with Parkinson's disease: Current knowledge and future prospects. *Front. Aging Neurosci.* **10**, 1 (2018).

49. Van Wamelen, D. *et al.* Exploring hyperhidrosis and related thermoregulatory symptoms as a possible clinical identifier for the dysautonomic subtype of Parkinson's disease. *J. Neurol.* **266**(7), 1736-1742 (2019).

50. Scott, B. M. *et al.* Cooccurrence of apathy and impulse control disorders in Parkinson disease. *Neurol.* **95**(20), 2769-2780 (2020).

51. Ffytche, D. H. *et al.* The psychosis spectrum in Parkinson disease. *Nat. Rev. Neurol.* **13**(2), 81-95 (2017).

52. Lenka, A., Pagonabarraga, J., Pal, P. K., Bejr-Kasem, H. & Kulisvesky, J. Minor hallucinations in Parkinson disease: A subtle symptom with major clinical implications. *Neurol.* **93**(6), 259-266 (2019).

53. Schapira, A. H. V., Chaudhuri, K. R. & Jenner P. Non-motor features of Parkinson disease. *Nat. Rev. Neurosci.* **18**(7), 435-450 (2017).

Author contributions statement

F.R.-S., C.R.-B., C.B., P.L., and P.M.-M. conceived the study. F.R.-S., C.B., and P.L. developed the methodology. F.R.-S. conducted the data analysis. All authors reviewed the results of the analysis and discussed the findings. F.R.-S. wrote the original draft of the manuscript. All authors reviewed the manuscript.

Funding

F.R.-S., C.B., and P.L. are supported in part by the Spanish Ministry of Economy and Competitiveness through the PID2019-109247GB-I00 project, by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3), and by the BBVA Foundation (2019 Call) through the "Score-based nonstationary temporal Bayesian networks. Applications in climate and neuroscience" project.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/srep

Data availability: All data, code and results are publicly available in our GitHub repository (https://github.com/ferjorosa/parkinson-subtypes).

Competing financial interests: The authors declare no competing financial interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformation.pdf