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

Fernando Rodriguez-Sánchez1, Carmen Rodriguez-Blazquez2,3, Concha Bielza1, Pedro Larrañaga1, Daniel Weintraub4, Pablo Martinez-Martin3, Alexandra Rizos5, Anette Schrag6 & K. Ray Chaudhuri5

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 psychosis, 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, Hoehn & Yahr stage, and treatment. Independent confirmation of these results could have implications for the clinical management of Parkinson's disease patients.

Parkinson's disease (PD) is a progressive neurodegenerative disease that is clinically characterized by a broad spectrum of motor and non-motor manifestations. 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).

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. However, to the best of our knowledge, all of them have used single-partition clustering methods such as k-means, latent class analysis, Gaussian mixture model, agglomerative hierarchical clustering. Single-partition clustering algorithms assume the existence of

---

1Computational Intelligence Group, Universidad Politécnica de Madrid, Madrid, Spain. 2National Center of Epidemiology, Carlos III Institute of Health, Madrid, Spain. 3Center for Networked Biomedical Research in Neurodegenerative Diseases (CIBERNED), Carlos III Institute of Health, Madrid, Spain. 4Departments of Psychiatry and Neurology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA. 5King’s College London, Department of Neurosciences, Institute of Psychiatry, Psychology & Neuroscience and Parkinson’s Foundation Centre of Excellence, King’s College Hospital, London, UK. 6Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London, UK. 7Email: fernando.rodriguezs@upm.es
work with missing information, no patients were excluded from the analysis. For more information about the
missing, mostly in the Sexual domain of the MDS-NMS. As our multi-partition clustering method was able to
for model-based multi-partition clustering, which extends model-based clustering 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-
partitions 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 Dis-
order 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. 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 addi-
tion, 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. For all patients,
socio-demographic 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), which is com-
posed 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 (maxi-
mum intensity). The total score of each part is obtained by summing the respective item scores.

2. The MDS-NMS, which is composed of 52 items grouped into 13 domains: depression, anxiety, apathy, psy-
chosis, 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, which ranges from 1 to 5.

Motor items from the MDS-UPDRS were classed as 5 motor cardinal signs: tremor, rigidity, bradykinesia, dys-
kinesias and motor fluctuations; and 2 motor subtypes: axial symptoms and postural instability gait difficulty
(PIGD). This resulted in 7 motor variables. The specific MDS-UPDRS items that constitute each motor vari-
able 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. Regarding
medication, 87% of the patients took levodopa. The average levodopa daily dose (LDD) for these patients was
658.57 ± 503.55 milligrams (mg). In addition, 42% of the patients received dopamine agonist (DA) treatment.
The average levodopa-equivalent daily dose of DA (LEDD-DA), calculated following Tomlinson et al., for these
patients was 226.84 ± 132.14 mg. Finally, 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.

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) 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 version of the structural expectation-maximization 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.

### 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 $\mu$ was used to denote the mean of this subtype for a specific symptom and the symbol $\sigma$ 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) moderate [0.25, 0.75], (iii) severe [0.75, 1].

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

| Variable | N | Mean (s.d.) | Median | Min | Max |
|----------|---|-------------|--------|-----|-----|
| **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) | 0.11   | 0.00 | 0.57 |
| Rigidity  | 398 | 0.19 (0.16) | 0.15   | 0.00 | 1.00 |
| Dyskinesias | 402 | 0.07 (0.15) | 0.00   | 0.00 | 1.00 |
| Motor Fluctuations | 401 | 0.16 (0.19) | 0.06   | 0.00 | 0.81 |
| Bradykinesia | 394 | 0.29 (0.17) | 0.25   | 0.00 | 0.89 |
| Axial symptoms | 402 | 0.23 (0.15) | 0.21   | 0.00 | 0.86 |
| PIGD      | 393 | 0.22 (0.19) | 0.15   | 0.00 | 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 |
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, PD duration, LDD, and LEDD-DA, 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, the presence of levo-dopa and DA treatments, and discrete variables such as the H&Y stage, χ² 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. In addition, treatment information of each subtype is provided in Table 3. Significant dif-
Table 2. Sex, age, age of onset, disease duration and H&Y stage of each PD subtype. Numbers between parentheses correspond to standard deviations. *Significant differences between A1 and A2. †Significant differences between B1 and B2. ‡Significant differences between C1 and C2. §Significant differences between C1 and C3. ¶Significant differences between D1 and D2. ‡‡Significant differences between E1 and E2. ‡‡‡Significant differences between F1 and F2. ‡‡‡‡Significant differences between H1 and H2. Statistical significance was defined as p-value p < 0.01.

Table 3. Treatment information of each subtype. Numbers between parentheses correspond to standard deviations. *Significant differences between A1 and A2. †Significant differences between B1 and B2. ‡Significant differences between C1 and C2. §Significant differences between C1 and C3. ¶Significant differences between C2 and C3. ‡‡Significant differences between D1 and D2. ‡‡‡Significant differences between E1 and E2. ‡‡‡‡Significant differences between F1 and F2. ‡‡‡‡‡Significant differences between H1 and H2. Statistical significance was defined as p-value p < 0.01.
ferences 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 (83%) was characterized by 335 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 (17%) was characterized by 67 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 199 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 ($\mu = 0.16, \sigma = 0.06$), and slight urinary issues ($\mu = 0.06, \sigma = 0.09$).
- Subtype B2 (51%) was formed of 203 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 ($\mu = 0.25, \sigma = 0.22$).

**Partition C (dyskinesias-psychosis).**

- Subtype C1 (60%) was composed of 242 patients that showed no dyskinesias ($\mu = 0.00, \sigma = 0.01$) or psychosis ($\mu = 0.00, \sigma = 0.00$).
- Subtype C2 (28%) was composed of 112 patients that showed slight dyskinesias ($\mu = 0.18, \sigma = 0.18$) and slight psychosis ($\mu = 0.02, \sigma = 0.03$).
- Subtype C3 (12%) was composed of 48 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 (67%) consisted of 272 patients that showed no mental fatigue ($\mu = 0.00, \sigma = 0.01$) and slight physical fatigue ($\mu = 0.14, \sigma = 0.06$).
- Subtype D2 (33%) consisted of 130 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 (49%) was constituted by 199 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 (51%) was constituted by 203 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 (53%) was composed of 214 patients that showed no orthostatic hypotension ($\mu = 0.00, \sigma = 0.02$) and slight sexual problems ($\mu = 0.01, \sigma = 0.03$).
- Subtype F2 (47%) was composed of 188 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 (79%) was characterized by 317 patients that showed no weight loss ($\mu = 0.00, \sigma = 0.01$) and slight depression ($\mu = 0.03, \sigma = 0.04$).
- Subtype G2 (21%) was characterized by 85 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 299 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 103 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 model10, the Gaussian mixture model11, and the unsupervised k-dependence Bayesian classifier19), and two model-based multi-partition clustering methods (i.e., the Gaussian expansion simplification until termination algorithm17, and the multi-partition mixture model18). 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 population21.

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 A2 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 observed25. Both socio-demographic aspects are known risk factors for ICRDs along with motor complications, a pre-PD history of ICRDs, and a DA treatment25. Related to this, we observed a higher percentage of DA treated patients in A2 than in A1. However, no causal relationship could be extracted from this observation.

Apathy, cognition, pain, gastrointestinal, sleep, and urinary symptoms were associated in partition B. Two subtypes were identified, where patients characterized as subtype B2 presented a higher severity of these symptoms. This subtype is consistent with the Parkinson’s apathy subtype34,35, which has been described to be formed by identifying a specific partition for fatigue symptoms, and connecting it with partitions (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, E2 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 phenotype26,30, a recent study has noted that normosmic PD patients usually display better motor function than hyposmic PD patients30.

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 F2 were male, showing significant differences in the sex of the patients belonging to F1 and F2. 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 dysfunction27–29.

Weight loss and depression were associated in partition G. Two subtypes were identified, where G2 was characterized by patients with mild weight loss and depression. Loss of appetite due to depression is a known weight loss factor30. There were no significant differences in sex, age, age of onset or H&Y stage of the patients.
belonging to G1 and G2. With regard to treatment, we did observe significant differences between subtypes. More specifically, there were considerably fewer number of patients with DA treatment in G2 than in G1, and those medicated patients were taking a significantly lower LEDD-DA.

Regarding clustering, anxiety was associated with excessive sweating. Anxiety was present in both the H1 and H2 subtypes, but patients in H2 showed a higher severity of this symptom along with a mild degree of excessive sweating. This result is consistent with a recent study 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 (A2 subtype) had a 0.75 probability of presenting the symptoms of B2. This result is consistent with a recent study that has challenged the traditional concept of apathy and ICRDs as opposite symptoms; (ii) patients that suffered psychosis (C3 subtype) had an 0.88 probability of suffering the symptoms of B2. The presence of visual hallucinations has been linked to sleep deprivation, cognitive impairment and depression; and (iii) patients with mild mental and physical fatigue (D2 subtype) had a 0.79 probability of suffering the symptoms of B2 and a 0.73 probability of suffering the symptoms of E2. As previously discussed, fatigue has been related to the presence of apathetic symptoms, sleep disturbances, and higher H&Y stages.

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.

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. Concerning MDS-UPDRS and MDS-NMS, these scales do not consider patient treatment. It is therefore difficult to identify if symptom severity is a natural consequence of PD or if it is a consequence of medication. Moreover, the majority of patients in this study were medicated. 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.

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

Received: 19 August 2021; Accepted: 29 November 2021
Published online: 08 December 2021

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 Roodeen, 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. Mi, 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. (Houghton & Mifflin, 1968).
11. McLachlan, G. J., Lee, S. X. & Rathnayake, S. I. Finite mixture models.
12. Gordon, A. D. A review of hierarchical classification. J. R. Stat. Soc. (Ser. 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(3), 1130 (2019).
14. Mestre, T. A. et al. Parkinson’s disease subtypes: critical appraisal and recommendations. J. Parkinsons Dis. 11(2), 395–404 (2021).
Author contributions

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.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-03118-w.

Correspondence and requests for materials should be addressed to F.R.-S.

Reprints and permissions information is available at www.nature.com/reprints.

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

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

© The Author(s) 2021