




OPEN

## Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories

Colin Birkenbihl<sup>3,4</sup>, Ashar Ahmad<sup>1,6,7</sup>, Nathalie J. Massat<sup>1,2,7</sup>, Tamara Raschka<sup>3,4</sup>, Andreja Avbersek<sup>1,5</sup>, Patrick Downey<sup>1</sup>, Martin Armstrong<sup>1</sup> & Holger Fröhlich<sup>3,4</sup>

Parkinson's disease (PD) is a highly heterogeneous disease both with respect to arising symptoms and its progression over time. This hampers the design of disease modifying trials for PD as treatments which would potentially show efficacy in specific patient subgroups could be considered ineffective in a heterogeneous trial cohort. Establishing clusters of PD patients based on their progression patterns could help to disentangle the exhibited heterogeneity, highlight clinical differences among patient subgroups, and identify the biological pathways and molecular players which underlie the evident differences. Further, stratification of patients into clusters with distinct progression patterns could help to recruit more homogeneous trial cohorts. In the present work, we applied an artificial intelligence-based algorithm to model and cluster longitudinal PD progression trajectories from the Parkinson's Progression Markers Initiative. Using a combination of six clinical outcome scores covering both motor and non-motor symptoms, we were able to identify specific clusters of PD that showed significantly different patterns of PD progression. The inclusion of genetic variants and biomarker data allowed us to associate the established progression clusters with distinct biological mechanisms, such as perturbations in vesicle transport or neuroprotection. Furthermore, we found that patients of identified progression clusters showed significant differences in their responsiveness to symptomatic treatment. Taken together, our work contributes to a better understanding of the heterogeneity encountered when examining and treating patients with PD, and points towards potential biological pathways and genes that could underlie those differences.

Parkinson's disease (PD) is an age-associated neurodegenerative disorder that affects approximately seven million people worldwide. Alongside the cardinal motor symptoms of bradykinesia, rigidity, resting tremor, and postural instability in later stages<sup>1</sup>, PD patients suffer from a wide range of non-motor symptoms such as sleep disturbances, psychosis, cognitive impairment, and mood disorders<sup>2</sup>. Currently there are no disease modifying treatments available for PD and present medications (e.g., L-DOPA) only offer symptomatic benefits. Designing and conducting clinical trials to test putative disease-modifying treatments is complicated due to the high inter-individual variability of disease progression rates<sup>3-5</sup>. Therefore, understanding the different biological mechanisms that drive differential disease progression is vital to ultimately pave the way for personalised therapies and can help to identify novel target candidates for therapeutic intervention.

Previous attempts to identify PD subtypes focused on ad-hoc classification of the motor characteristics of tremor (tremor dominant sub-type) and postural instability (postural instability and gait dominant sub-type)<sup>6</sup>. Similarly, age at disease diagnosis has been used to classify PD patients into Late Onset PD and Young Onset PD<sup>3</sup>. However, given the broad and complex range of PD symptoms, single-variable subtyping approaches are unlikely to capture the complexity of patients' progression. Here, data-driven multivariate approaches using, for example, cluster analysis<sup>5</sup> offer a promising opportunity to overcome these limitations.

The foundation for such multivariate subtyping approaches is built through multi-modal longitudinal data provided by observational cohort studies such as the Parkinson's Progression Markers Initiative (PPMI)<sup>7</sup>. PPMI

<sup>1</sup>UCB Pharma, Chemin du Foriest 1, 1420 Braine-L'Alleud, Belgium. <sup>2</sup>Veramed Limited, 5th Floor Regal House, 70 London Road, Twickenham TW1 3QS, UK. <sup>3</sup>Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany. <sup>4</sup>Bonn-, Aachen International Center for IT, University of Bonn, Friedrich Hirzebruch-Allee 6, 53115 Bonn, Germany. <sup>5</sup>Present address: Regeneron Inc., 777 Old Saw Mill River Road, Tarrytown, NY 10591, USA. <sup>6</sup>Present address: Grünenthal GmbH, 52078 Aachen, Germany. <sup>7</sup>These authors contributed equally: Ashar Ahmad and Nathalie J. Massat. ✉email: colin.birkenbihl@scai.fraunhofer.de

data has been previously used to identify patient subtypes based on cross-sectional imaging data and cerebrospinal fluid biomarkers at study baseline<sup>2,8</sup>. Only a few studies have focused on disease progression which requires the use of longitudinal follow-up data. This aspect was partially addressed by Faghri et al.<sup>9</sup> using PPMI data at 48 months follow-up. The authors identified three PD subtypes using non-negative matrix factorisation. Still, their approach was unable to discern these subtypes with respect to the slope of progression. In this context, recently published neural network-based approaches make it possible to cluster entire longitudinal patient trajectories<sup>10,11</sup>. However, these studies did not explore the biological underpinning of the subtypes nor did they consider how their patients differed in their clinical presentation or in their response to treatment.

The aim of this work was to uncover PD progression clusters by applying an artificial intelligence-based, purely data-driven approach based on multivariate longitudinal trajectories comprised of motor and non-motor scores obtained from *de-novo* patients. Furthermore, using machine learning, we sought to identify associations linking discovered progression clusters to potentially disparate biological pathways, genetic variations, and clinical symptoms. Finally, we aimed to assess any difference in the loss of dopaminergic neurons across clusters and whether patients of distinct progression clusters would respond differently to symptomatic treatment. Such insights could contribute to a deeper understanding and characterisation of the heterogeneous mechanisms at play within PD and offer the opportunity to define novel drug targets.

## Results

**Multivariate time series analysis identifies three patient clusters with distinct progression profiles.** By clustering the time series data of 407 *de novo* PD patients from PPMI (267 male, 140 female) using our previously published artificial intelligence-based VaDER approach<sup>11</sup>, we identified three groups of PD patients with distinct progression profiles (Supplementary Section S1, Fig. S1). The clustering was conducted based on the multivariate progression of six key clinical assessments of PD symptoms over the course of up to 60 months: the MDS-UPDRS 1, 2, and 3 (off treatment)<sup>12</sup>, tremor dominant score (TD), postural instability and gait disorder score (PIGD), and the Epworth sleepiness scale (ESS).

The three resulting clusters contained 'moderate'-progressors (n = 230), 'fast'-progressors (n = 53), and 'slow'-progressors (n = 124). Table 1 provides summary statistics of patients from each cluster at study baseline. We found significant differences between the average age at study baseline of slow progressors and the two other respective subtypes (t-test 'slow' versus 'fast',  $p < 0.013$ ; 'slow' versus 'moderate',  $p < 0.019$ ; 'moderate' versus 'fast',  $p > 0.32$ ). In contrast, no significant difference was observed in the elapsed time from initial diagnosis to study baseline (pairwise U-tests between all three clusters,  $p > 0.3$ ), or distribution of Hoehn and Yahr stages ( $\chi^2$ -test,  $p > 0.15$ ). With respect to MDS-UPDRS scores at study baseline, we found a significant difference in MDS-UPDRS 1 between the 'moderate' cluster and the other two clusters, respectively (U-test, 'slow' versus 'fast',  $p < 0.01$ ; 'moderate' versus 'fast',  $p < 0.001$ ; 'slow' versus 'moderate',  $p > 0.59$ ). For MDS-UPDRS 2, the only significant deviation was observed comparing the 'moderate' against 'fast'-progressors (U-test, 'moderate' versus 'fast',  $p < 0.025$ ; 'slow' versus 'fast',  $p > 0.14$ ; 'slow' versus 'moderate',  $p > 0.34$ ). We identified no significant difference in MDS-UPDRS 3 scores (pairwise U-test for all clusters,  $p > 0.69$ ). Furthermore, we detected no significant differences in the distribution of biological sex ( $\chi^2$ -test,  $p > 0.15$ ) and the start of symptomatic therapy (Fig. S2).

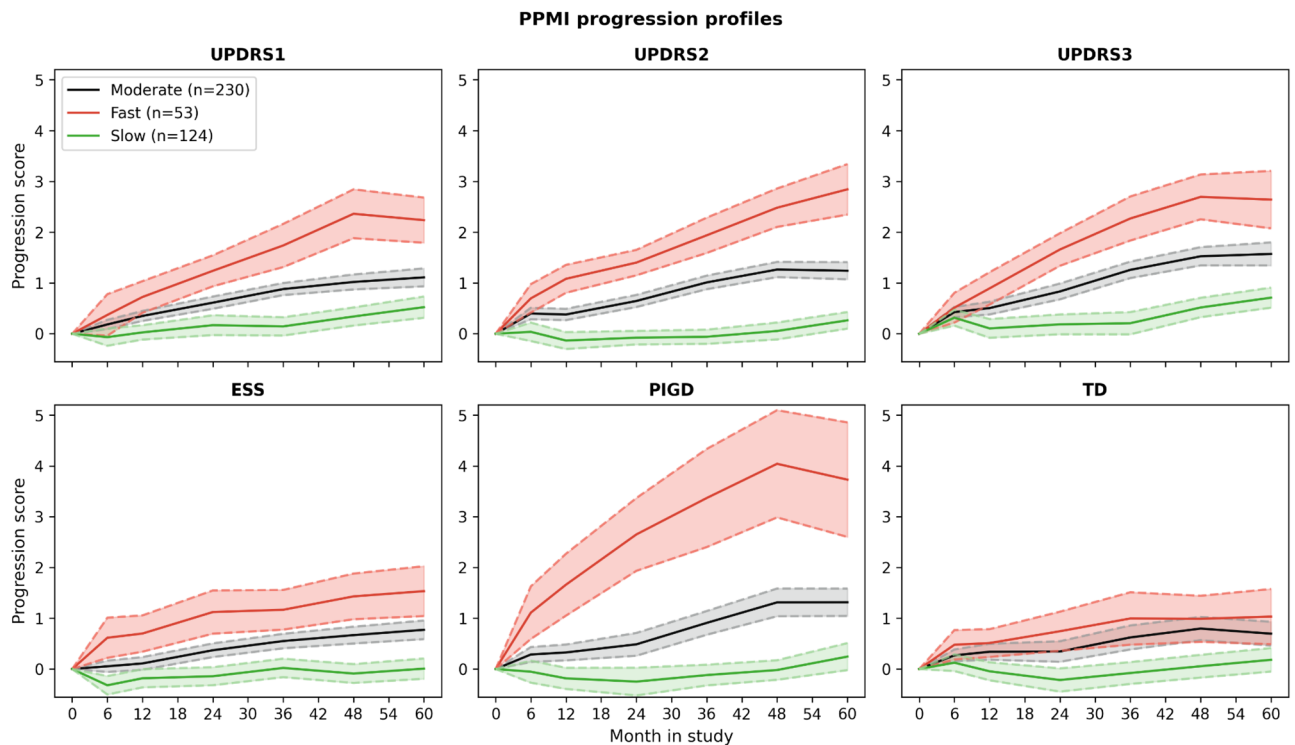
The mean univariate progression trajectories of these clusters along with their 95% confidence intervals are depicted in Fig. 1. Although the clustering was conducted on multiple outcome measures, we observed a clear separation of clusters across all selected variables except for the TD score between 'fast' and 'moderate' progressors. While 'fast' and 'moderately' progressing subtypes displayed a clear increase of symptoms over the covered 60 month interval already starting from baseline, 'slow'-progressors experienced almost no significant symptom worsening across scores until month 24.

## Characterisation of PD clusters suggests longitudinal differences in dopaminergic deficiency.

The differences in motor symptom progression rates across subtypes (Fig. 1) were mirrored by significant differences in the age-adjusted trajectories of DaTSCAN measurements, which were available until month 48: the rate in loss of specific-binding ratio (SBR) signal in the caudate region was significantly lower for the cluster exhibiting 'slow' progression than for both the 'fast' and 'moderate' progressing clusters, respectively (signal loss of  $-0.0033$  SBR unit/month, 95% CI [ $-0.0055$ ,  $-0.0011$ ],  $p = 0.004$  compared to the 'fast' group, and of  $-0.0019$  SBR unit/month, 95% CI [ $-0.0032$ ,  $-0.0003$ ],  $p = 0.01$  compared to the 'moderate' group). No significant difference in SBR was observed between the 'fast' and 'moderate' progressing groups (details in Supplementary Section S3). The difference in rate of dopaminergic loss between the 'fast' and the 'slow' progressing clusters was seen equally in the ipsilateral (signal loss of  $-0.0034$  SBR unit/month, 95% CI [ $-0.0056$ ,  $-0.0008$ ],  $p = 0.008$ ) and the contralateral (signal loss of  $-0.0032$  SBR unit/month, 95% CI [ $-0.0057$ ,  $-0.0008$ ],  $p = 0.007$ )

Cluster	N	Age (Years) *	Number of Females	Years since diagnosis	UPDRS 1*	UPDRS 2*	UPDRS 3	Hoehn and Yahr stage I	Hoehn and Yahr stage II	Hoehn and Yahr stage III
Slow	124	60.2 ± 9.3	49 (40%)	0.5 ± 0.5	5.4 ± 4.2	5.9 ± 4.2	21.0 ± 9.1	57 (46%)	65 (52%)	2 (0.2%)
Moderate	230	62.7 ± 9.7	78 (34%)	0.6 ± 0.5	5.1 ± 3.5	5.4 ± 3.9	20.6 ± 8.7	95 (41%)	135 (59%)	0 (0%)
Fast	53	64.2 ± 10.8	13 (26%)	0.7 ± 0.8	7.3 ± 4.7	7.2 ± 4.9	21.0 ± 8.8	27 (51%)	26 (49%)	0 (0%)

**Table 1.** Summary statistics of patients per subtype at study baseline. UPDRS refers to the MDS-UPDRS scale. Presented is the mean and standard deviation of variables as well as the percentage of females per subtype. N, Number of patients per subtype. \*Differences were statistically significant;  $p$ -values are provided in the Result section.



**Figure 1.** Mean trajectories of the three different progression clusters. Dashed lines depict the 95% confidence interval of the respective trajectory. Confidence intervals grow larger with time as more patients drop-out of the study. The progression score depicted on the y-axis represents the relative change to study baseline normalised by the standard deviation of the respective variable. UPDRS refers to the MDS-UPDRS testing battery, ESS to the Epworth Sleepiness Scale, PIGD to the Postural Instability Gait Disorder, and TD to the Tremor Dominant Score.

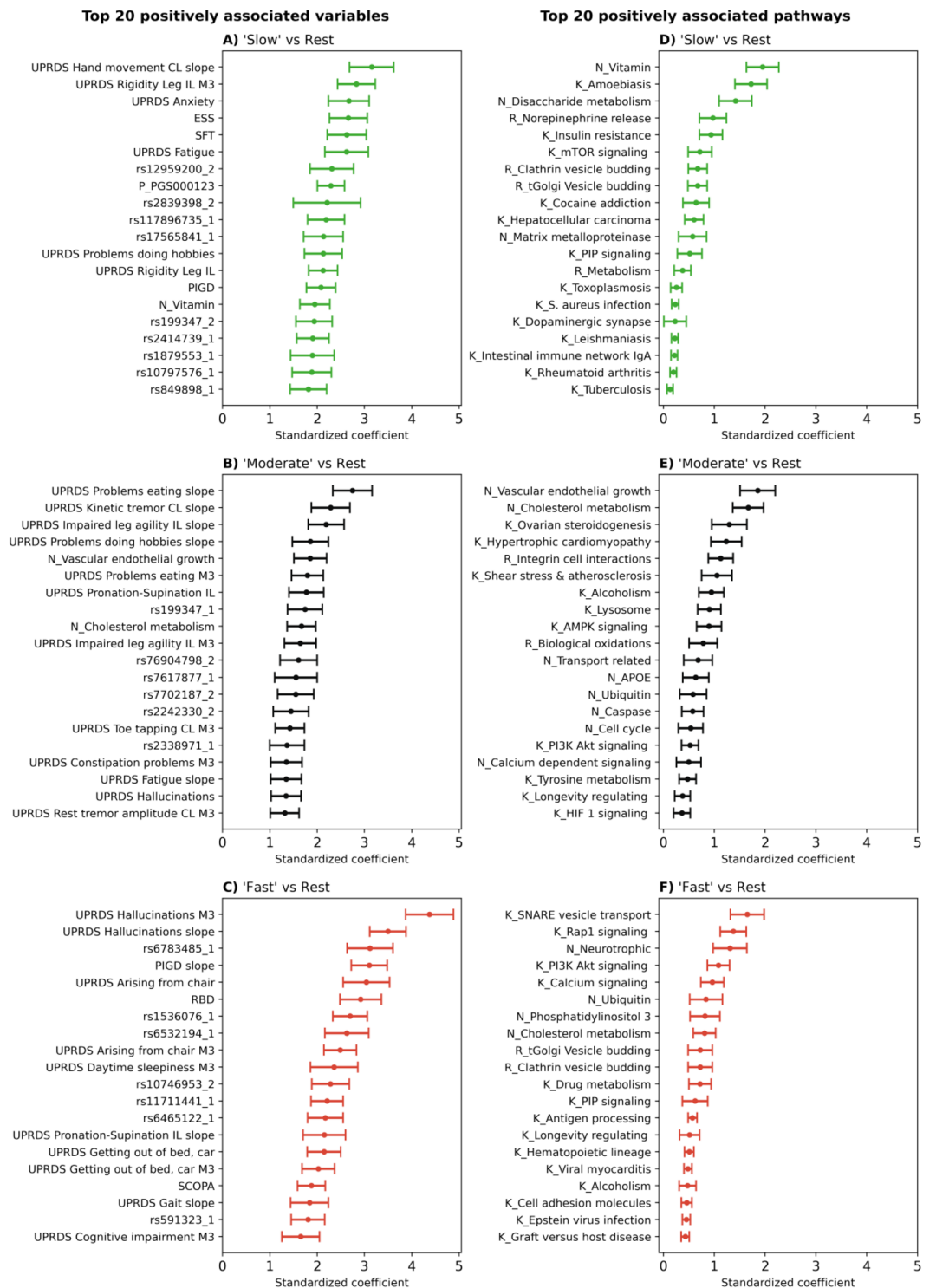
sides of the caudate region. In contrast, the difference in rate of progression between the 'moderate' and the 'slow' progressing subtypes was stronger in the contralateral side (signal loss of  $-0.0022$  SBR unit/month, 95% CI [ $-0.0038, -0.0006$ ],  $p=0.006$ ) as compared to the ipsilateral side (signal loss of  $-0.0016$  SBR unit/month, 95% CI [ $-0.0030, +0.0002$ ],  $p=0.07$ ) sides of the caudate region. No significant difference in SBR rates were observed in the putamen, and changes in the striatum were intermediary between those observed in the caudate and the putamen.

**Machine learning revealed associations between clusters and underlying biology.** To discover further associations between the identified progression clusters and clinical as well as biomarker and genetic variables, we developed machine learning models based on patients' baseline visit data. Additionally, we built a second version of these models that included the 3-month follow-up data, both in the form of raw values and of change relative to baseline values. The variables included into the models comprised demographic and clinical data, including MDS-UPDRS item-level data (86 variables at baseline; 217 including 3 month follow-up), CSF biomarkers (amyloid beta, phosphorylated tau, total tau), blood serum transcriptomic data (7 variables), 3472 SNPs gained through a linkage disequilibrium analysis of an initial set of 145 PD associated SNPs obtained from DisGeNET<sup>13</sup>, and brain region specific DaTSCAN (5 variables). We also calculated burden-scores for biological pathways stemming from Kegg<sup>14</sup>, Reactome<sup>15</sup>, and NeuroMMSig<sup>16</sup> (36, 10, and 12 pathways, respectively). These scores were based on the SNP data of each respective patient and described the amount of genetic variation affecting a pathway (see Method section for details). A full list of all variables is presented in the Supplementary Spreadsheet.

The machine learning algorithm of choice was a sparse group LASSO (SGL)<sup>17</sup>. We developed three distinct models, each discriminating one of the clusters from the respective other two (i.e., one versus rest approach). The significance of the most strongly associated variables was then determined by bootstrapping each model 200 times and investigating whether the resulting confidence intervals (CI) of standardised coefficients contained zero. CIs were Bonferroni-corrected to account for multiple testing. Further methodological details are described in Supplementary Section S4.

The built models revealed several significant associations between measured variables and progression clusters, which were interpretable from a clinical as well as a biological point of view.

**Progression clusters are associated with distinct symptoms and genetic loci.** The coefficients of each machine learning model highlight how specific variables influence the probability that a patient belongs



**Figure 2.** Top 20 variables associated with the respective progression cluster (sparse group LASSO using baseline data + 3-month follow-up). The plots show the standardised coefficient together with their Bonferroni-corrected 95% confidence intervals for each variable. A stronger positive coefficient value in the plot indicates a higher likelihood of a patient belonging to the respective cluster. A corresponding plot for baseline data only is shown in Fig. S7. (A–C, most associated variables for ‘slow’, ‘moderate’ and ‘fast’ progression. The number after SNP IDs indicates the number of non-reference alleles. ‘M3’ denotes variables measured at the 3 month visit. ‘slope’ indicates the calculated slope of the corresponding score measured 3 months after baseline. PGS denotes polygenic risk scores. ‘CL’ means contralateral, while ‘IL’ refers to ipsilateral. (D–F), most associated biological pathways. Pathways starting with ‘K\_’, ‘R\_’, or ‘N\_’ originate from Kegg, Reactome, and NeuroMMSig, respectively.

to a particular cluster. For interpretability, we focused on significant positive interactions (i.e., variables that increase the chance of belonging to the respective cluster; Fig. 2A–C).

The variable most strongly associated with ‘fast’ PD progression was the presence and severity of hallucinations at the 3 month follow-up visit (NP1HALL m3, 95%CI [3.91, 5.0]), with the increase in experienced hallucinations following in third position (NP1HALL slope, 95%CI [3.07, 3.9]). In fourth position, the increase in postural instability and gait disorder severity over the first 3 months was found (PIGD slope, 95% CI [2.73, 3.55]). Additionally, ‘fast’ progressing patients experienced more difficulties when rising from a lying or sitting position compared to the other two subtypes (95% CI: NP3RISNG [2.56, 3.63], NP3RISNG m3 [2.16, 2.98], NP2RISE m3 [1.9, 2.65], NP2RISE [1.8, 2.64]). REM sleep behaviour disorder (RBD) proved to be another association for ‘fast’ progression (95% CI [2.33, 3.24]). Furthermore, several SNPs (rs6783485-LOC105377110, rs1536076-SH3GL2, rs6532194-chromosome 4:89859751, rs11711441–chromosome 3:183103487, and rs591323-LOC105379297) were found to be among the top 20 associated variables for ‘fast’ progression. Notably, all these SNPs were taken from DisGeNET, because of their known association to PD according to GWAS studies. In all cases, the non-reference-allele increased the risk of ‘faster’ PD progression.

‘Slow’ PD progression was associated with increasing difficulties when performing the hand movement task of the MDS-UPDRS (NP3HMOV slope 95% CI [2.93, 3.38]). Furthermore, a series of highly associated variables were connected to daytime sleepiness (ESS 95% CI [2.27, 3.06]) and general fatigue (NP1FATG 95% CI [2.16, 2.97]). Patients of the ‘slow’ cluster also suffered more often from anxiety (95% CI: NP1ANXS [2.15, 2.93]; NP1ANXS m3 [0.89, 1.53]) and were the only subtype which showed a significant positive association with depression, albeit the coefficient remained rather small (geriatric depression scale 95% CI [0.1, 0.65]). Additionally, better semantic fluency was also connected to ‘slower’ disease progression (SFT 95% CI [2.06, 2.84]). With regard to motor symptoms, ‘slow’ progression was associated with rigidity of the ipsilateral extremities at baseline, month 3, and their relative increase in severity (95% CI: NP3RIGL\_IL m3 [2.23, 3.09]; NP3RIGL\_IL [1.74, 2.54]; NP3RIGU\_IL [1.0, 1.61]). Further, we found a significant positive association of the polygenic risk score PGS000123<sup>18</sup> and multiple genetic loci with the probability to belong to the ‘slow’-progressors. SNPs rs17565841 (OCA2), and rs12959200 (chromosome 18:73599819) placed among the top 10 associations (95% CI: [2.11, 2.71], [1.95, 3.05], [1.91, 2.77], respectively). Once again, these SNPs were taken from DisGeNET because of their known association to PD according to GWAS studies.

For ‘moderate’ disease progression, the strongest association was the worsening of performing the eating task of the MDS-UPDRS over the first 3 months (NP2EAT slope 95% CI [2.3, 3.08]). Further, reduced agility in the ipsilateral leg was associated with ‘moderate’ progression (95% CI: NP3LGAG\_IL slope [1.79, 2.55]; NP3LGAG\_IL m3 [1.36, 2.06]). With rs76904798 (chromosome 12:40220632), rs199347 (GPNMB), rs7702187 (SEMA5A), and rs7617877 (LINC00693), we identified several PD associated SNPs which raised the probability for patients to belong to the ‘moderate’ subtype.

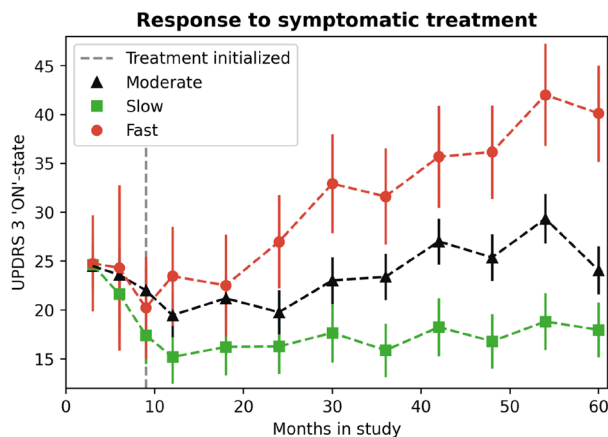
A comprehensive view on all variables and their coefficients can be found in the Supplementary Spreadsheet.

While the SGLs were designed to identify variable associations and not to make reliable forecasts, we additionally evaluated their predictive performance. With a cross-validated area under the receiver operating characteristic curve of 0.62, 0.60, and 0.63 for ‘slow’, ‘moderate’, and ‘fast’ progression, respectively, their performance remained limited.

**Genetic burden scores connect the heterogeneity in PD progression to biological pathways.** Several biological pathways and genes could be associated with the respective clusters (Fig. 2 D–F). The ‘fast’ cluster was highly associated with higher genetic burden in the Kegg ‘SNARE vesicle transport’ pathway (95% CI [1.25, 1.92]), the ‘Rap1 signalling’ pathway (95% CI [1.1, 1.71]), and NeuroMMSig’s ‘neurotrophic’ subgraph (95% CI [1.25, 1.92]). The patients of the ‘moderate’ cluster were linked to the ‘cholesterol metabolism’ subgraph (95% CI [1.56, 2.25]) and ‘vascular endothelial growth factor’ subgraph (95% CI [1.42, 2.12]) originating from NeuroMMSig. The ‘vitamin’ and ‘disaccharide metabolism’ subgraphs from NeuroMMSig, and Kegg’s ‘amoebiasis pathway’ were discovered as strongly associated with the ‘slow’ progressing clusters (95% CI: [1.6, 2.22], [1.04, 1.66], and [1.14, 1.86], respectively). A list of all mappings between pathways, genes and SNPs can be found in the Supplementary Spreadsheet.

**Identified clusters show differences in response to motor symptom therapy.** After observing that potentially different biological pathways were involved in the PD pathology of each cluster, we investigated whether the clusters also differed in their response to symptomatic treatment for motor symptoms. To this aim, we selected participants who had initiated Levodopa or Dopamine agonist symptomatic treatment between month 6 and month 9 after baseline and assessed whether progression as measured by MDS-UPDRS 3 score differed by PD cluster. We separately analysed the ‘ON’-state MDS-UPDRS 3 score data, in which patients are examined approximately one hour after taking medication (Fig. 3), and the ‘OFF’-state MDS-UPDRS 3 score data (Fig. S11). As per PPMI protocol, patients were considered to be in the ‘OFF’-state when the last treatment dose was taken at least 6 h before symptoms were assessed<sup>19</sup>. Methodological details can be found in Supplementary Section S6.

Although initially all three PD clusters responded similarly to symptomatic treatment by stabilising their motor scores in the first 9 months after treatment initiation (i.e. 9–18 months post-baseline, Fig. 3, Fig.S11), we observed that patients in the ‘fast’ progressing cluster continued to progress fastest and all three clusters had significantly different MDS-UPDRS 3 scores in ‘ON’ and ‘OFF’-states at 30 months after baseline (i.e. 21 months post-symptomatic treatment initiation) from each others, i.e. the 95% CIs did not overlap. PD subtypes did not differ according to whether they were prescribed Levodopa (alone or in combination with Dopamine agonist), or Dopamine agonist alone as a first line of PD symptomatic treatment (Table S1). The levodopa equivalent daily



**Figure 3.** Differential response to symptomatic treatment. Effect plot of modelled MDS-UPDRS 3 'ON'-state score progression prior to and after the initiation of Levodopa or Dopamine agonist in patients who initiated therapy between 6 and 9 months post-baseline using a longitudinal LMEM with time fitted as a categorical variable and baseline score fitted as a covariate. The error bars represent the 95% confidence intervals, based on standard errors computed from the covariance matrix of the fitted regression coefficients.

dose (LEDD) was obtained for the PPMI participants included in this analysis (Table S2). Only beyond 42 months post-baseline, patients in the 'fast' cluster appeared to have taken higher LEDD compared to the patients in the 'moderate' cluster (mean difference at month 54: 186.8, 95%CI [76.2, 267.6],  $p < 0.01$ ), while no significant difference was found for 'fast' versus 'slow', and 'slow' versus 'moderate' progressors, respectively (Figure S12).

## Discussion

In this work, we identified three distinct PD progression clusters dividing patients into 'slow', 'moderate', and 'fast'-progressors. This clustering built on the multivariate trajectory of six clinical variables rather than a single univariate outcome. Investigation of potential confounders that could have biased the clustering showed no significant differences of biological sex, disease duration, and Hoehn & Yahr stages across clusters. Also with respect to the type of symptomatic treatment and LEDD, no bias was identified in our clustering. A machine learning model further identified significant associations between clinical measurements taken at study baseline (optionally including 3 months follow-up data), genetic features, biological pathways, and the different progression clusters of patients. Several distinct SNPs and biological mechanisms could be associated with each cluster. Analysis of the observed associations provided insights into the heterogeneity of PD progression and the distinct biological pathways potentially promoting it. Further analysis revealed that patients in different clusters responded differently to symptomatic treatment and displayed significant differences in dopaminergic cell loss. Altogether this makes it improbable that our clustering is just a consequence of patients being in different disease stages at study baseline.

Our clustering differentiates itself from previous clustering approaches in various ways: 1) instead of relying on snapshot, cross-sectional data at any arbitrary point in time, we focus on the progression of key clinical variables over time, 2) this progression is modelled multivariate to better represent the natural progression of PD which occurs across multiple scales, 3) through the inclusion of pathway-specific genetic perturbation scores, we can generate hypotheses connected to possible differences in PD pathology across the identified clusters, 4) we analysed the difference in symptomatic treatment response across clusters, which was seldomly done before<sup>20</sup>.

**Interpretation of significant associations between variables and PD progression clusters.** Our machine learning models identified that measurements taken early in the disease course already show significant associations with the longitudinal progression of PD's motor and non-motor symptoms. Such significant associations, however, do not imply that the majority of patients in a respective cluster experienced a strongly associated symptom, instead, it indicates that patients suffering from that specific symptom are statistically more likely to belong to the associated cluster. Further, while we found statistically significant differences in MDS-UPDRS 1 & 2 total scores and items between 'faster' progressing patients and the other two clusters, we identified significant associations of individual non-motor symptoms measured via the MDS-UPDRS items with every cluster. This highlights the importance of going beyond high-level clinical assessments when investigating symptom manifestation across PD subgroups.

In the 'fast'-progressing clusters, the presence of psychotic symptoms in the form of hallucinations or delusions was found as the strongest association. Indeed, hallucinations can already be observed in newly diagnosed patients<sup>21</sup> and experiencing such visual or auditory hallucinations was established to be one of the most notable risk factors for increased mortality<sup>22</sup> and earlier placement in care homes<sup>23</sup>. These findings could, on the one hand, be explained by the difficulties of living with psychosis but, on the other, also point towards a faster disease progression in general. In this context, the association between RBD and our 'fast' progressing cluster is noteworthy, as RBD is one of the major risk factors for hallucinations<sup>24</sup> and was also hypothesised to be an early

sign of faster disease progression<sup>25</sup>. Furthermore, RBD has been connected to reduced striatal dopaminergic activity<sup>26</sup>, which is in line with our observations for the ‘fast’ progressing cluster. In concordance, Wang et al. discovered slower and faster progressing subtypes based on brain pathology with the faster subtype showing increased RBD and decreased dopaminergic brain efficiency in the caudate and putamen at study baseline<sup>27</sup>. In another subtyping effort by Fereshtehnejad et al. a ‘diffuse malignant’ PD cluster was described that showed faster disease progression and was characterised by lower CSF amyloid beta values<sup>28</sup>. Indeed, our ‘fast’ progressing cluster was also associated with lower amyloid beta in CSF, however, considerably older and more affected by hallucinations than the presented ‘diffuse malignant’ subtype. Since the investigated PPMI patients were de novo PD patients, the significant difference in age across clusters at baseline added further evidence to a previously discovered trend that patients with later disease onset often experience faster progression<sup>29,30</sup>.

The ‘slow’ cluster showed strong associations with non-motor symptoms such as fatigue, sleepiness, and anxiety. While these symptoms have received increasing recognition in recent years, they remain poorly understood aspects of PD<sup>2</sup> and little is known about disease progression in patients that suffer from them. Previously, a more benign PD progression was noticed among patients with resting tremor<sup>31</sup>, a finding that was in concordance with our analysis that linked ‘slow’ progression to resting tremor as measured through MDS-UPDRS item 3.17.

Previous case series reported on several associations between slower disease progression and attributes we found to be significant associations with what we called ‘moderate’ progression<sup>32</sup>. Here, it was described that patients with predominantly worsening tremors, younger age, and no indication of PGID showed reduced disease progression.

Only slight differences in global cognitive performance as measured by the Montreal Cognitive Assessment (MoCA) could be found among the clusters. This could be due to the comparably early time point of assessment (approximately one month after PD diagnosis for most patients), since only subtle cognitive changes are observable in the PPMI cohort over the first 5 years<sup>2</sup>. However, semantic fluency was among the strongest associated variables with ‘slow’ progression, indicating that this cluster could be more stable with respect to cognitive performance. Patients who suffered from cognitive symptoms measured by the MDS-UPDRS were most often encountered in the ‘fast’ progressing cluster.

The limited predictive performance of the SGLs can be explained by the relatively small sample sizes of the identified clusters, the modelling strategy which was primarily chosen to identify significant associations rather than to provide predictions, as well as the difficulty of predicting PD progression from baseline measures. Previous attempts on predicting future PD progression based on baseline variables also reported limited performance in external validation<sup>30</sup>.

**PD progression clusters are associated with distinct biological pathways and gene mutation load.** With the inclusion of available genetic data into the models, we were able to identify distinct biological pathways that were associated with the different clusters. This opens up the opportunity of not only identifying new therapeutic targets, but targets that may be positioned more effectively within certain subgroups of patients.

The pathway most predominantly associated with ‘fast’-progression was the Kegg ‘SNARE interactions in vesicular transport’ pathway. Vesicle dysfunction is a known phenomenon in the pathogenesis of PD, the targeting of related proteins (including SCNA and LRRK2) has been discussed for several years now<sup>33</sup> and there are multiple lines of supporting evidence for the role of this pathway in PD. In this pathway, the retrieved SNPs predominantly mapped to genes encoding for vesicle associated membrane proteins (VAMP2, VAMP4) and syntaxins (SXT4, and SXT1B). VAMP2 interacts with SXT1 in the neuronal synapse and is important for vesicle fusion and neurotransmitter translocation<sup>34,35</sup>. VAMP4 and syntaxins interact with LRRK2<sup>36</sup>, a major PD risk factor and potential drug target in which mutations promote a PD phenotype<sup>37</sup>, with respect to retrograde and post-Golgi signalling. Both VAMP2 and SXT1 showed diagnostic potential in blood-based biomarker studies for PD<sup>38</sup>.

The second strongest association found for fast progressors was the ‘Rap1 signaling’ pathway which is involved in the nigrostriatal dopaminergic pathway in medium spiny neurons<sup>39</sup>. Again, ample evidence lends biological support to the role of this pathway, including the position of the vascular endothelial growth factor (VEGFA) gene in the pathway, that has been shown to protect dopaminergic neurons from cell death. VEGFA has been discussed as a potential target for treating PD<sup>40</sup> and a recent study suggests blocking of VEGFA to prevent blood–brain-barrier disruption, which has been implicated in several neurodegenerative diseases, including PD<sup>41</sup>.

Furthermore, this pathway involves several fibroblast growth factors (FGF5, 10, and 20), with FGF20 also being a prominent entity in the ‘Neurotrophin’ mechanism listed in NeuroMMSig (the third most associated pathway for ‘fast’ progression). The FGF gene family has also been associated with neuroprotection and neurogenesis, partially by triggering PI3K-AKT signalling which also occurred among our highly associated pathways with respect to ‘fast’ PD progression<sup>42</sup>.

Taken together, it can be postulated that severe perturbations in Golgi vesicle transport that eventually cause apoptosis, in combination with a reduced neuroprotection and neurogenesis to replace damaged cells might promote a ‘fast’ progressing form of PD.

The ‘moderately’ progressing cluster was mainly associated with NeuroMMSig’s ‘Vascular endothelial growth factor’ and ‘Cholesterol metabolism’ pathways. The former was largely defined by VEGFA which was discussed above and might indicate a common mechanism between ‘fast’ and ‘moderate’ progressors. The squalene synthase (FDFT1) was the major gene in the ‘cholesterol metabolism’ pathway to which we could map SNPs. Squalene is an antioxidant and precursor of cholesterol which is essential for synaptic functioning and has been linked to PD and  $\alpha$ -synuclein aggregation<sup>43</sup>. This, along with additional supporting evidence for this pathway<sup>44–49</sup>, could indicate that oxidative stress might play a more pronounced role in ‘moderately’ progressing PD compared to the other two subtypes.

The strongest associated pathway for the ‘slow’ progressing cluster was the ‘Vitamin subgraph’ which evolved around the solute carrier family 41 member 1 (SLC41A1). This gene is part of the PD related PARK16 locus and is associated with magnesium efflux and homeostasis which is believed to contribute to PD<sup>50</sup>. Furthermore, the ‘amoebiasis’ pathway was identified as the second highest associated and the connection of the underlying genes to PD has been observed previously<sup>51</sup>. Interestingly, we also found an association of ‘slow’ progressing PD to the ‘disaccharide metabolism’ pathway, in which GBA was a key agent. Whilst GBA mutation carriers were not included in the analysed sporadic PD PPMI cohort, three SNPs in our analysis could still be mapped to GBA, (rs2230288, rs12752133, and rs76763715) and all have been associated to an increased risk of PD<sup>52</sup>.

**Differential response to symptomatic motor treatment across progression clusters.** When the progression of motor symptoms was compared between the clusters after the initiation of Levodopa and/or Dopamine agonists, a substantial difference in the response to the symptomatic treatment was observed, which could not be explained either by medication dosage or type of therapy. Together with the observed genetic differences between clusters, our results strongly suggest that the identified progression clusters represent an inherent property of the disease. Notably, differential response to symptomatic treatment for PPMI de-novo PD cohort participants with fastest motor progression was also reported in<sup>53</sup>, and by Lawton et al. using data from the Tracking Parkinson and Oxford Parkinson's Disease Centre Discovery cohort<sup>54</sup>.

**Limitations.** When interpreting the genetic data, it should be noted that our SNP inclusion was hypothesis driven based on prior evidence for an association with PD. Nevertheless, the work presented highlights the ability of the models to discriminate between molecular pathways involved in the different clusters, and the importance of genetic data in PD. The availability of larger datasets with attached genome wide genetic data would support a more hypothesis generating approach and potentially uncover novel mechanisms. Further, our approach relies on a clinical diagnosis of PD. While the PD diagnosis of patients was repeatedly confirmed over the several year long follow-up of PPMI, a potential misdiagnosis of patients could bias the results and the retention time of patients in the prodromal phase of PD remains unknown. Finally, PPMI as a primary data source for our analysis is an observational study in which patients are treated according to best clinical routine practice. The treatment itself is not monitored precisely, thus, the entirety of medication taken by patients, their treatment compliance, as well as a potential presence of residual medication effects remain unknown. The minimum 6 h medication washout defined by PPMI might be too short when extended release formulations were administered to patients. However, as the LEDD calculation takes into account the type of formulation of the dopaminergic therapy, as well as the impact of any adjuvant therapy, it is unlikely that this biased our clustering as no significant difference in either the type of medication nor the LEDD was observed across clusters.

## Conclusion

Using our clustering approach, we show that PD patients can be divided into ‘slow’, ‘moderate’, and ‘fast’-progressors based on the relative change of symptoms over the time course of the study. These groups not only show differences in the progression rates of clinical symptoms but also differ in the rate of dopaminergic cell loss, and importantly respond differently to symptomatic treatment. An analysis of whole genome sequencing data also suggests that genetic and mechanistic differences underpin these groupings. Currently, several agents are being tested in the clinic for their ability to slow disease progression but running such trials in a group of patients containing individuals with very different progression rates is fraught with difficulty. In the PPMI cohort that we used in this work, we identified 124 of 407 patients as slow progressors, and these patients showed no worsening of any symptom for at least 24 months. Given that current disease modifying trials in PD do not exceed two years, one can expect about a third of the patients to show no symptom worsening for the duration of the trial, provided that PPMI can be regarded as a representative PD study. As disease modifying treatments do not aim to improve symptoms but to slow down their worsening then the presence of a significant number of slow progressors who will not deteriorate during the trial will make it very difficult to observe disease slowing in a mixed population even with a highly effective treatment.

Future work is needed to further validate our established PD progression clusters ideally with the help of a larger study where similar data modalities as in PPMI are measured in de-novo PD patients.

## Materials and methods

**Dataset and patient selection criteria.** We selected 407 de-novo PD patients from the PPMI dataset. Our inclusion criteria were: age older than 30 years, Hoehn and Yahr stage of 1 or 2, recent PD diagnosis, and untreated by anti-PD medication (patient in the off-state according to the PPMI data). Furthermore, we used only patients with at least 48 months of follow-up. PPMI acquired informed consent to data collection and sharing from all participating individuals and got ethical approval. Ethical guidelines on human data collection were adhered to.

**Preprocessing by calculating progression scores.** To enable a cluster of patients along their disease progression, we transformed the selected variables into ‘progression scores’ that capture each variable’s change relative to baseline. We calculated these progression scores by subtracting the baseline value from the value measured at each respective time point and dividing the result by the variables standard deviation at baseline. When training the machine learning models, the raw baseline (or month three) measurements were taken and standardised or one-hot-encoded (ie., in contrast to the clustering they were progression agnostic).



**Multivariate clustering of clinical trajectories.** Optimal hyperparameters for the VaDER model were found following the procedure described in<sup>11</sup>: We evaluated several possible models using a varying set of hyperparameters (including the number of sought clusters) and, finally, selected the hyperparameters which led to the best model performance. The performance of the model was quantified by comparing the prediction strength of the model against a random subtyping of the same data. We selected the smallest number of clusters that showed a significant difference to a random clustering with respect to the achieved prediction performance (Fig. S1). The clustering was repeated 20 times and the final subtypes were assigned based on a consensus clustering across the 20 repeats. Supplementary Section S1 provides further details, including diagnostic plots.

**Characterisation of PD progression clusters.** *Analysis of dopaminergic deficiency.* DaTSCAN data were analysed for differences between PD clusters over time. Data from baseline up to 48 months was considered. Participants without DaTSCAN screening data (N = 17) were excluded from the analysis, leaving data for 390 participants. The longitudinal progression profile for individual patients in each cluster is shown in Fig. S6. Details about the statistical analysis are presented in Supplementary Section S3.

*Response to symptomatic therapy.* Patients were defined as being on symptomatic treatment, if they were taking L-DOPA, or dopamine agonists, with or without other types of motor symptom therapy such as MAO-B inhibitors at a respective visit<sup>19</sup>. Since a relatively highest fraction of patients started treatment at 9 months of follow-up, we focused our analysis on this time point. Altogether 44 in the 'slow' cluster started a symptomatic treatment at 9 months, 67 in the 'moderate', and 16 patients in the 'fast' cluster. The longitudinal progression profile using loess smoothing for individual patients in each cluster is shown in Fig. S8. Details about the statistical analysis including diagnostic plots are presented in Supplementary Section S6.

*Analysis of whole genome sequencing data.* PPMI provides whole genome sequencing (WGS) data of de novo diagnosed PD patients. To reduce the extreme high dimensionality of the WGS data while taking into account the very limited sample size, we focused only on single nucleotide polymorphisms (SNPs) with putative association to PD. More specifically, we obtained an initial list of 646 PD associated SNPs obtained from GWAS Catalogue<sup>55</sup>, PheWas<sup>56</sup>, and DisGeNET<sup>13</sup>. This list was subsequently expanded via linkage disequilibrium analysis (LD,  $r^2 > 0.8$ ) using Haploreg<sup>57</sup>, which also provides a gene mapping based on proximity. In addition, we employed a cis-eQTL mapping via GTex<sup>58</sup> to associate SNPs to genes expressed in brain tissues. Altogether 14520 SNPs were mapped to 1055 genes. In a second step, the genes were further mapped onto 12 PD specific mechanisms defined in the NeuroMMSig database<sup>16</sup>, as well as 36 KEGG<sup>14</sup> and 10 Reactome<sup>15</sup> pathways that were significantly enriched for PD associated genes. How we calculated the pathway scores based on the selected SNPs is presented in the Supplementary Section S4.

## Data availability

The authors have no permission to directly share any of the patient-level data as stated by the data usage agreement with the original data owners (the PPMI study). The PPMI data used in this work can be accessed at [www.ppmi-info.org](http://www.ppmi-info.org) after successful access application.

Received: 2 August 2022; Accepted: 14 February 2023

Published online: 18 February 2023

## References

1. Postuma, R. B. *et al.* MDS clinical diagnostic criteria for parkinson's disease. *Mov. Disord.* **30**(12), 1591–1601 (2015).
2. Weintraub, D. & Mamikonyan, E. The neuropsychiatry of parkinson disease: A perfect storm. *Am. J. Geriatr. Psychiatry* **27**(9), 998–1018 (2019).
3. Thenganatt, M. A. & Jankovic, J. Parkinson disease subtypes. *JAMA Neurol.* **71**(4), 499–504 (2014).
4. Sieber, B. A. *et al.* Prioritized research recommendations from the National Institute of Neurological Disorders and Stroke Parkinson's Disease 2014 conference. *Annals of neurology* **76**(4), 469–472 (2014).
5. Van Rooden, S. M. *et al.* The identification of parkinson's disease subtypes using cluster analysis: A systematic review. *Mov. Disord.* **25**(8), 969–978 (2010).
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. Marek, K. *et al.* The parkinson progression marker initiative (PPMI). *Prog. Neurobiol.* **95**(4), 629–635 (2011).
8. Erro, R. *et al.* Clinical clusters and dopaminergic dysfunction in de-novo parkinson disease. *Parkinsonism Relat. Disord.* **28**, 137–140 (2016).
9. Faghri, F., Hashemi, S. H., Leonard, H., Scholz, S. W., Campbell, R. H., Nalls, M. A., & Singleton, A. B. Predicting onset, progression, and clinical subtypes of parkinson disease using machine learning. *bioRxiv*, 338913 (2018).
10. Zhang, X. *et al.* Data-driven subtyping of parkinson's disease using longitudinal clinical records: A cohort study. *Sci. Rep.* **9**(1), 1–12 (2019).
11. de Jong, J. *et al.* Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience* **8**(11), giz134 (2019).
12. 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. Off. J. Mov. Disord. Soc.* **23**(15), 2129–2170 (2008).
13. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**(D1), D845–D855 (2020).
14. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**(suppl\_1), D480–D484 (2007).
15. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**(D1), D498–D503 (2020).
16. Domingo-Fernández, D. *et al.* Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): A web server for mechanism enrichment. *Bioinformatics* **33**(22), 3679–3681 (2017).

17. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **22**(2), 231–245. <https://doi.org/10.1080/10618600.2012.681250> (2013).
18. Ibanez, L. *et al.* Parkinson disease polygenic risk score is associated with parkinson disease status and age at onset but not with alpha-synuclein cerebrospinal fluid levels. *BMC Neurol.* **17**(1), 1–9 (2017).
19. Simuni, T. *et al.* Longitudinal change of clinical and biological measures in early parkinson's disease: Parkinson's progression markers initiative cohort. *Mov. Dis.* **33**(5), 771–782 (2018).
20. Marras, C. & Lang, A. Parkinson's disease subtypes: Lost in translation?. *J. Neurol. Neurosurg. Psychiatry* **84**(4), 409–415 (2013).
21. Pagonabarraga, J. *et al.* Minor hallucinations occur in drug-naive parkinson's disease patients, even from the premotor phase. *Mov. Disord.* **31**(1), 45–52 (2016).
22. Weil, R. S. & Reeves, S. Hallucinations in parkinson's disease: New insights into mechanisms and treatments. *Adv. Clin. Neurosci. Rehabil. ACNR* **19**(4), 189 (2020).
23. Goetz, C. G. & Stebbins, G. T. Risk factors for nursing home placement in advanced parkinson's disease. *Neurology* **43**(11), 2222–2222 (1993).
24. Pacchetti, C. *et al.* Relationship between hallucinations, delusions, and rapid eye movement sleep behavior disorder in parkinson's disease. *Mov. Dis. Off. J. Mov. Disord. Soc.* **20**(11), 1439–1448 (2005).
25. Fereshtehnejad, S. M. *et al.* New clinical subtypes of parkinson disease and their longitudinal progression: A prospective cohort comparison with other phenotypes. *JAMA Neurol.* **72**(8), 863–873 (2015).
26. Eisenstein, I. *et al.* Reduced striatal dopamine transporters in idiopathic rapid eye movement sleep behaviour disorder: Comparison with parkinson's disease and controls. *Brain* **123**(6), 1155–1160 (2000).
27. Wang, L. *et al.* Association of specific biotypes in patients with parkinson disease and disease progression. *Neurology* **95**(11), e1445–e1460 (2020).
28. Fereshtehnejad, S. M., Zeighami, Y., Dagher, A. & Postuma, R. B. Clinical criteria for subtyping parkinson's disease: Biomarkers and longitudinal progression. *Brain* **140**(7), 1959–1976 (2017).
29. Maetzler, W., Liepelt, I. & Berg, D. Progression of parkinson's disease in the clinical phase: Potential markers. *Lancet Neurol.* **8**(12), 1158–1171 (2009).
30. Latourelle, J. C. *et al.* Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed parkinson's disease: A longitudinal cohort study and validation. *Lancet Neurol.* **16**(11), 908–916 (2017).
31. Josephs, K. A., Matsumoto, J. Y. & Ahlskog, J. E. Benign tremulous parkinsonism. *Arch. Neurol.* **63**(3), 354–357 (2006).
32. Foltynie, T., Brayne, C. & Barker, R. A. The heterogeneity of idiopathic parkinson's disease. *J. Neurol.* **249**(2), 138–145 (2002).
33. Oeda, T. *et al.* Impact of glucocerebrosidase mutations on motor and nonmotor complications in parkinson's disease. *Neurobiol. Aging* **36**(12), 3306–3313 (2015).
34. Bittner, M. A. & Holz, R. W. Kinetic analysis of secretion from permeabilized adrenal chromaffin cells reveals distinct components. *J. Biol. Chem.* **267**(23), 16219–16225 (1992).
35. Schoch, S. *et al.* SNARE function analyzed in synaptobrevin/VAMP knockout mice. *Science* **294**(5544), 1117–1122 (2001).
36. Beilina, A. *et al.* The parkinson's disease protein LRRK2 interacts with the GARP complex to promote retrograde transport to the trans-golgi network. *Cell Reports* **31**(5), 107614 (2020).
37. Cookson, M. R. The role of leucine-rich repeat kinase 2 (LRRK2) in parkinson's disease. *Nat. Rev. Neurosci.* **11**(12), 791–797 (2010).
38. Agliardi, C. *et al.* Oligomeric  $\alpha$ -Syn and SNARE complex proteins in peripheral extracellular vesicles of neural origin are biomarkers for parkinson's disease. *Neurobiol. Dis.* **148**, 105185 (2021).
39. Zhang, X. *et al.* Balance between dopamine and adenosine signals regulates the PKA/Rap1 pathway in striatal medium spiny neurons. *Neurochem. Int.* **122**, 8–18 (2019).
40. Axelsen, T. M. & Woldbye, D. P. Gene therapy for parkinson's disease, an update. *J. Parkinsons Dis.* **8**(2), 195–215 (2018).
41. Ebanks, K., Lewis, P. A. & Bandopadhyay, R. Vesicular dysfunction and the pathogenesis of parkinson's disease: Clues from genetic studies. *Front. Neurosci.* **13**, 1381 (2019).
42. Liu, Y., Deng, J., Liu, Y., Li, W. & Nie, X. FGF, mechanism of action, role in parkinson's disease, and therapeutics. *Front. Pharmacol.* **12**, 1572 (2021).
43. García-Sanz, P., MFGAerts, J. & Moratalla, R. The role of cholesterol in  $\alpha$ -synuclein and lewy body pathology in gba1 parkinson's disease. *Mov. Dis.* **36**(5), 1070–1085 (2021).
44. Huang, X. *et al.* Serum cholesterol and the progression of parkinson's disease: Results from DATATOP. *PLoS One* **6**(8), e22854 (2011).
45. Sere, Y. Y., Regnacq, M., Colas, J. & Berges, T. A *Saccharomyces cerevisiae* strain unable to store neutral lipids is tolerant to oxidative stress induced by  $\alpha$ -synuclein. *Free Radical Biol. Med.* **49**(11), 1755–1764 (2010).
46. Kabuto, H., Yamanushi, T. T., Janjua, N., Takayama, F. & Mankura, M. Effects of squalene/squalane on dopamine levels, antioxidant enzyme activity, and fatty acid composition in the striatum of parkinson's disease mouse model. *J. Oleo Sci.* **62**(1), 21–28 (2013).
47. Sánchez-Pernaute, R. *et al.* Selective COX-2 inhibition prevents progressive dopamine neuron degeneration in a rat model of parkinson's disease. *J. Neuroinflammation* **1**(1), 1–11 (2004).
48. Van't Erve, T. J. *et al.* Reinterpreting the best biomarker of oxidative stress: The 8-iso-prostaglandin F2 $\alpha$ /prostaglandin F2 $\alpha$  ratio shows complex origins of lipid peroxidation biomarkers in animal models. *Free Radical Biol. Med.* **95**, 65–73 (2016).
49. Onodera, Y., Teramura, T., Takehara, T., Shigi, K. & Fukuda, K. Reactive oxygen species induce Cox-2 expression via TAK1 activation in synovial fibroblast cells. *FEBS Open Bio.* **5**, 492–501 (2015).
50. Sturgeon, M., Perry, W. & Cornall, R. SLC41A1 and TRPM7 in magnesium homeostasis and genetic risk for parkinson's disease. *J. Neurol. Neurosurg.* **1**(9), 23 (2016).
51. Wang, J., Liu, Y. & Chen, T. Identification of key genes and pathways in parkinson's disease through integrated analysis. *Mol. Med. Rep.* **16**(4), 3769–3776 (2017).
52. Huang, Y., Deng, L., Zhong, Y. & Yi, M. The association between E326K of GBA and the risk of parkinson's disease. *Parkinsons Dis.* **2018**, 1048084 (2018).
53. Tsiouris, K. M., Konitsiotis, S., Koutsouris, D. D. & Fotiadis, D. I. Prognostic factors of Rapid symptoms progression in patients with newly diagnosed parkinson's disease. *Artif. Intell. Med.* **103**, 101807 (2020).
54. Lawton, M. *et al.* Developing and validating parkinson's disease subtypes and their motor and cognitive progression. *J. Neurol. Neurosurg. Psychiatry* **89**(12), 1279–1287 (2018).
55. Buniello, A. *et al.* The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**(D1), D1005–D1012 (2019).
56. Denny, J. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1111. <https://doi.org/10.1038/nbt.2749> (2013).
57. Ward, L. D. & Kellis, M. HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**(D1), D877–81. <https://doi.org/10.1093/nar/gkv1340> (2016).
58. GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**(6), 580–585. <https://doi.org/10.1038/ng.2653> (2013).

## Acknowledgements

This project was partially funded by the EU-wide ERAPerMed project DIGIPD (01KU2110) and the European Union's Horizon 2020 research and innovation program under grant agreement No. 826421, "TheVirtualBrain-Cloud". Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database [www.ppmi-info.org/data](http://www.ppmi-info.org/data). PPMI—a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners. A list of names of all the PPMI funding partners can be found at [www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors/](http://www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors/). This work was partially performed, while A.Ah. and H.F. were affiliated with UCB Monheim and A.Av. was affiliated with UCB Braine.

## Author contributions

Designed the project: P.D., M.A., H.F.; supervised the project: M.A., H.F., A.Av., P.D.; analysed the data and implemented algorithms: C.B., T.R., N.J.M., A.Ah.; drafted the manuscript: C.B., H.F., M.A., P.D., A.Av., N.J.M.; all authors have read and approved the manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

PD and MA are employees of UCB BioPharma. HF, AAh, and AAv were full time employees of UCB BioPharma at the start of this study. NJM is a Veramed statistical consultant for UCB Biopharma.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-30038-8>.

**Correspondence** and requests for materials should be addressed to C.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons 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) 2023