npj | parkinson's disease

Article

Published in partnership with the Parkinson's Foundation

https://doi.org/10.1038/s41531-025-00983-4

Predicting dementia in people with Parkinson's disease

Mohamed Aborageh[1] ✉, Tom Hähnel[1,2], Patricia Martins Conde[3], Jochen Klucken[3,4] & Holger Fröhlich[1,5] ✉

Parkinson's disease (PD) exhibits a variety of symptoms, with approximately 25% of patients experiencing mild cognitive impairment and 45% developing dementia within ten years of diagnosis. Predicting this progression and identifying its causes remains challenging. Our study utilizes machine learning and multimodal data from the UK Biobank to explore the predictability of Parkinson's dementia (PDD) post-diagnosis, further validated by data from the Parkinson's Progression Markers Initiative (PPMI) cohort. Using Shapley Additive Explanation (SHAP) and Bayesian Network structure learning, we analyzed interactions among genetic predisposition, comorbidities, lifestyle, and environmental factors. We concluded that genetic predisposition is the dominant factor, with significant influence from comorbidities. Additionally, we employed Mendelian randomization (MR) to establish potential causal links between hypertension, type 2 diabetes, and PDD, suggesting that managing blood pressure and glucose levels in Parkinson's patients may serve as a preventive strategy. This study identifies risk factors for PDD and proposes avenues for prevention.

Parkinson's disease (PD) is the fastest growing neurological disease and the second most common neurodegenerative disease[1]. Although the cause of PD remains unclear in most sporadic cases, mutations in more than 20 genes have been associated with the disease, including genes such as LRRK2 and SNCA[2]. In addition, environmental risk factors have been linked to the development of PD, which can be triggered by an underlying genetic predisposition[3]. For example, occupational exposures such as pesticides are associated with an increased risk of PD[4–6]. Furthermore, lifestyle factors such as alcohol consumption, physical inactivity and diet[7], as well as beta-blockers[8] and comorbidities such as diabetes mellitus and arterial hypertension, have been identified as potential risk factors.

Although PD is characterized primarily by three main motor symptoms (bradykinesia, tremor at rest, and rigidity); cognitive impairment is also a frequent problem in patients with PD (PwPD). Approximately 25% of PwPD exhibit mild cognitive impairment and 45% develop dementia within 10 years after diagnosis[9–14]. PD dementia (PDD) contributes to an increase in health-related expenditures and a significant decrease in quality of life, underscoring its importance to affected individuals, their caregivers and healthcare systems[15]. Furthermore, PDD constitutes one of the four milestones that occur on average four years before death, which indicates the onset of the terminal phase of the disease[16]. The underlying reason why certain PwPD develop cognitive impairment earlier in their disease course is still unknown. Previous studies report possible associations of different genetic risk variants with cognitive impairment in PD, including variants in SNCA[17,18], APOE[19,20], MAPT[21,22], TMEM175[23] and GBA[24,25]. Statistical analysis of a cohort of 827 individuals identified age, duration of the disease, sex, and GBA status as the primary factors associated with cognitive performance and progression to dementia[26]. Harvey et al. (2022) developed machine learning models to predict cognitive outcomes in de novo diagnosed PwPD using a broad spectrum of baseline data (including cognitive test results) from the Parkinson's Progression Markers Initiative (PPMI) cohort study[27]. However, data from specific cohorts are not necessarily representative of the overall disease population and could therefore differ from real-world data collected in population studies or clinical routine. Moreover, models trained on data collected solely for research, such as DNA methylation, cerebral spinal fluid biomarkers and neuropsychiatric tests, are not easily applicable in clinical practice. Thus, there is a need to investigate models using more routinely collected health data, including comorbidities, lifestyle and environmental factors. The potential impact of such factors on PDD is important for shaping future prevention strategies.

Our study used the UK Biobank (UKB) to a) predict dementia in people with Parkinson's disease (PwPD) using machine learning and b) investigate modifiable risk factors that can causally influence the

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53757 Sankt, Augustin, Germany. [2]Department of Neurology, University Hospital and Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany. [3]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg. [4]Centre Hospitalier de Luxembourg (CHL), Luxembourg, Luxembourg. [5]Bonn-Aachen International Center for Information Technology (B-IT), Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany. ✉e-mail: mohamed.aborageh@scai.fraunhofer.de; holger.froehlich@scai.fraunhofer.de

Table 1 | Demographic characteristics of Parkinson's disease (PD) and Parkinson's disease dementia (PDD) groups in the UK Biobank and PPMI datasets

Dataset	UK Biobank		PPMI	
Trait	**PD**	**PD dementia**	**PD**	**PD dementia**
Age	62.7 (5.3)	64.1 (4.4)	60.3 (9.9)	67.1 (7.9)
Sex				
Male	2172 (61.3%)	340 (69.8%)	310 (60.1%)	75 (61.5%)
Female	1369 (38.7%)	147 (30.2%)	205 (39.9%)	47 (38.5%)
Education				
Age at completion of full-time education	16.5 (2.4)	16.5 (2.4)	–	–
Years of education	–	–	15.7 (3.4)	14.2 (4.1)
Hoehn & Yahr				
H&Y I	–	–	141 (27.37%)	16 (13.11%)
H&Y II	–	–	337 (65.43%)	82 (67.21%)
H&Y III	–	–	22 (4.27%)	17 (13.93%)
H&Y IV	–	–	2 (0.38%)	2 (1.63%)
H&Y V	–	–	0 (0%)	3 (2.45%)

Continuous variables are reported as mean (standard deviation), while categorical variables are presented as count (percentage).

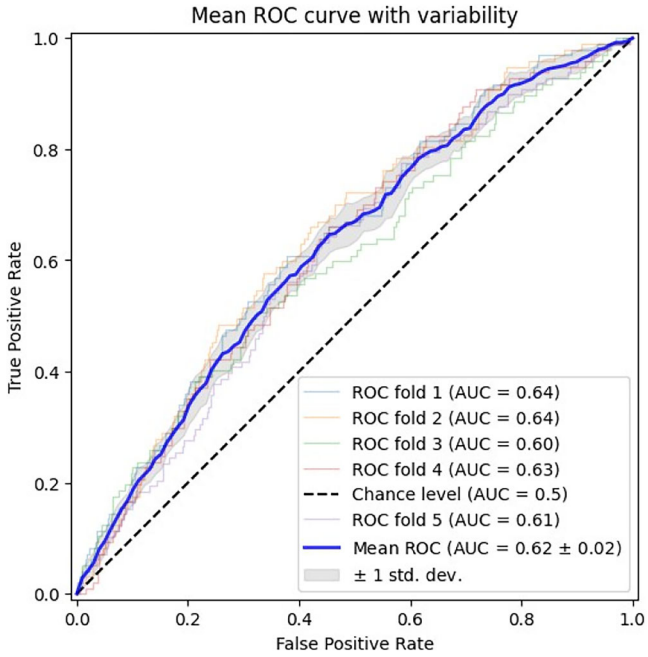

Fig. 2 | Mean ROC curve with variability. Mean ROC curve of the Random Forest model, derived from five-fold cross-validation. Individual fold ROC curves are shown as thin colored lines, and the bold blue line represents the mean ROC across folds. The shaded area denotes ±1 standard deviation, illustrating variability in performance. The diagonal dashed line indicates chance-level performance (AUC = 0.5). ROC receiver operating characteristic, AUC area under the curve.

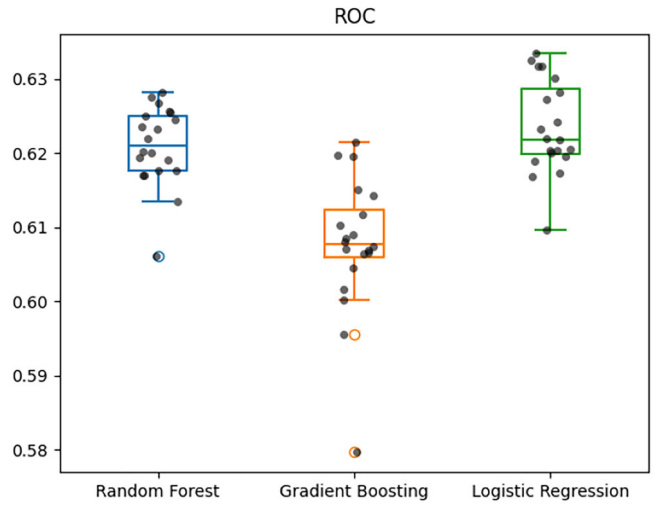

Fig. 1 | Boxplot of ROC scores across models. Boxplot showing the models' AUC via repeated nested cross-validation. The boxplots are displayed with a median line, the interquantile range (IQR) represented by the box borders, whiskers extending to 1.5 the IQR and outliers represented by hollow circles. Each black dot represents a repetition of the nested cross-validation of each model. ROC receiver operating characteristic, AUC area under the curve.

development of dementia in PwPD, thus informing possible prevention strategies. To our knowledge, no prior research has explored these areas.

To achieve our objectives, we developed and evaluated machine learning models to predict PDD based on comorbidities, genetic, environmental, and lifestyle factors. Using Explainable AI (XAI) techniques, we identified the features that significantly affected predictions and used probabilistic graphical models (Bayesian networks) to analyze interactions among the most predictive comorbidities, environmental, lifestyle, and genetic risk factors. Finally, we performed a Mendelian Randomization (MR) analysis to confirm the probable causal relationship between hypertension, type 2 diabetes, and PDD.

Results

Demographics of patients and controls

In UKB, PwPD were more likely to be male (61.3%) with a mean age of 62.7 years, consistent with previously reported incidences[28,29]. The PDD group showed an even higher proportion of males (69.8%) and a slightly higher mean age of 64.1 years.

In PPMI, PwPD showed a proportion of male individuals similar to UKB (60.4%) with a mean age of 61.8 years. The PDD group showed a slightly lower proportion of male individuals than in UKB (61.5%), and the mean age was 67.1 years. Detailed demographic characteristics of the groups can be found in Table 1.

Predictability of PD dementia

We evaluated penalized logistic regression, Random Forests, and XGBoost to predict PDD. Repetitive nested cross-validation showed an average AUC of 0.62 for Random Forest and logistic regression and 0.61 for XGBoost over 20 repetitions (Figs. 1 and 2). We carried out an ablation study to understand the contribution of different data modalities. We found that demographics, comorbidities, and genetics contributed the most, with a statistically significant average of 2% drop in AUC after removing any of them; see Tables 2 and 3.

We repeated our analysis using the PPMI dataset and compared the results with those obtained from the UKB cohort. Due to the absence of several variables in the PPMI dataset that were only available in UKB, we had to take a reduced subset of variables. This subset included only SNPs, PRS, age, and sex. When training the Random Forest model on UKB with this reduced subset of variables, the cross-validated prediction performance of this reduced model was close to the original with an average AUC 0.61 ± 0.01, while the cross-validated prediction performance on PPMI was higher with an average AUC 0.65 ± 0.02 (Fig. 3).

Table 2 | Comparison of the model's performance showing the median AUC over 20 repeats of nested cross-validation, using all data modalities and each modality separately

Modality	Random forest		XGBoost		Logistic regression	
	Median AUC	*P*-value	Median AUC	*P*-value	Median AUC	*P*-value
Multimodal	0.62 (0.005)	–	0.60 (0.009)	–	0.62 (0.006)	–
Demographics + Genetics	0.61 (0.004)	<0.05	0.59 (0.007)	<0.05	0.61 (0.006)	<0.05
Demographics	0.56 (0.005)	<0.05	0.54 (0.008)	<0.05	0.57 (0.004)	<0.05
Genetics	0.58 (0.003)	<0.05	0.56 (0.008)	<0.05	0.58 (0.05)	<0.05
Comorbidities	0.58 (0.004)	<0.05	0.57 (0.004)	<0.05	0.57 (0.03)	<0.05
Lifestyle	0.49 (0.01)	<0.05	0.5 (0.009)	<0.05	0.49 (0.005)	<0.05
Environmental	0.51 (0.007)	<0.05	0.51 (0.008)	<0.05	0.5 (0.006)	<0.05
Family history	0.49 (0.01)	<0.05	0.49 (0.01)	<0.05	0.49 (0.006)	<0.05

A Kruskal–Wallis test was used to determine if there were statistically significant differences in performance after the removal of each modality compared to the full model.

Table 3 | Ablation study results showing the median AUC over 20 repeats of nested cross-validation after removing each data modality

Ablation	Random forest		XGBoost		Logistic regression	
	Median AUC	*P*-value	Median AUC	*P*-value	Median AUC	*P*-value
Multimodal	0.62 (0.005)	–	0.60 (0.009)		0.62 (0.006)	
-Demographics	0.60 (0.004)	<0.05	0.58 (0.009)	<0.05	0.59 (0.005)	<0.05
-Genetics	0.59 (0.008)	<0.05	0.58 (0.009)	<0.05	0.60 (0.006)	<0.05
-Comorbidities	0.61 (0.005)	<0.05	0.59 (0.007)	<0.05	0.61 (0.005)	<0.05
-Lifestyle	0.62 (0.005)	0.19	0.61 (0.01)	0.07	0.63 (0.004)	<0.05
-Environmental	0.62 (0.005)	0.35	0.61 (0.009)	0.4	0.62 (0.006)	0.68
-Family history	0.62 (0.004)	0.35	0.61 (0.007)	0.74	0.62 (0.006)	0.78

A Kruskal–Wallis test was used to determine if there were statistically significant differences in performance after the removal of each modality relative to the full model.

SHAP analysis

For the following analysis, we focused on the best performing model variant after training and tuning on the entire dataset, i.e. the Random Forest model using all data modalities.

SHAP analysis revealed that variables from different data modalities contribute significantly to the predictions of the model, with the polygenic risk score PGS4281 being the main predictor, followed by SNP *rs769449*, age, SNP *rs6859*, diagnosis of depression, sex, BMI and hypercholesterolemia. We also observed that lifestyle variables, such as tea intake and water intake, appeared among the top predictors. Figure 4 shows the impact of the top 20 variables on the model output and Figure 5 shows the relationship between variable values and model predictions using SHAP dependence. We found that higher PGS4281 scores and older age were associated with higher SHAP values, indicating a stronger influence on the predicted probability of PDD. In contrast, a higher BMI was associated with lower SHAP values, i.e. lower predicted likelihood of PDD. For SNP *rs769449*, having one or two copies of the effect allele resulted in higher SHAP values. In contrast, for SNPs *rs6859* and *rs449647*, having more copies of the effect allele was associated with lower SHAP values. Furthermore, a higher predicted likelihood of PDD was associated with being male and having a diagnosis of depression.

Finally, we analyzed the cumulative influence of different data modalities by summing the SHAP values of their respective variables to understand their overall contribution to model predictions. In agreement with the previously presented ablation study, we found that genetics had the highest influence (49.31%), followed by demographics (24.32%) and comorbidities (15.74%). Figure 6 presents the cumulative influence of all data modalities.

The SHAP analysis of the Random Forest trained on the entire PPMI data set showed similar results with age as the main predictor, followed by PGS4281, SNP *rs2927468*, *rs449647* and *rs356219*, while sex and *rs769449* had lower SHAP values. A comparison of the beeswarm plots for both datasets, highlighting the SHAP values of the top predictors, is provided in Supplementary Fig. 1.

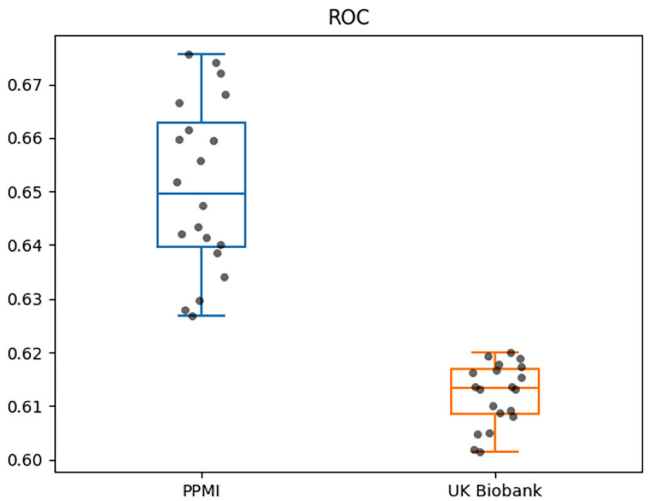

Fig. 3 | Boxplot of ROC scores on reduced subset of features. Comparison of the Random Forest model trained on the reduced subset of features in PPMI and UK Biobank datasets. The boxplots are displayed with a median line, the interquantile range (IQR) represented by the box borders, whiskers extending to 1.5 the IQR and outliers represented by hollow circles. Each black dot represents a repetition of the nested cross-validation of each model. ROC receiver operating characteristic, AUC area under the curve.

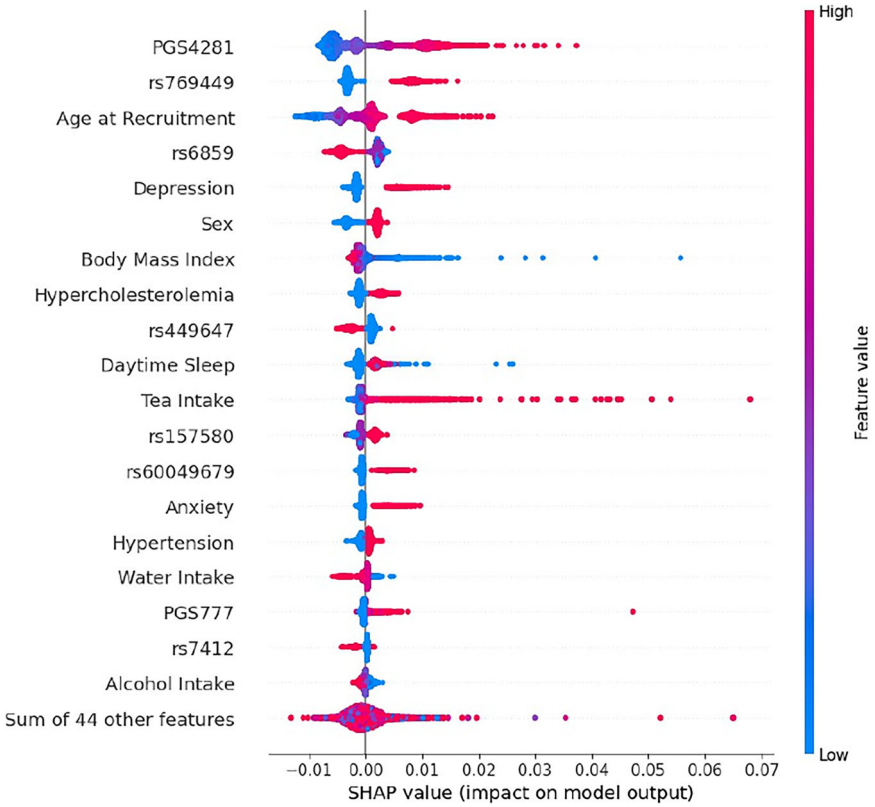

Fig. 4 | SHAP summary plot. Beeswarm plot showing the contribution of the top features to the model's output. Each point represents a single observation, with the horizontal axis indicating the SHAP value (impact on model output). Features are ranked by mean absolute SHAP value, from most to least important. Color represents the original feature value (red = high, blue = low), illustrating how feature magnitude influences model predictions. The bottom entry aggregates the combined impact of 44 additional features not shown individually.

Understanding the interactions of predictors

Following our SHAP analysis, we fitted 1000 BNs within a bootstrap procedure to better understand conditional statistical dependencies between all variables in the best-performing Random Forest model trained on UKB. All edges discussed below had a bootstrap frequency of (≥50%), which means that they were observed in at least 50% of the BNs across all bootstrap samples; thus, these edges had high statistical confidence. A visualization of these high-confidence edges as a graph is provided in Supplementary Fig. 2.

Overall, our analysis revealed expected connections within each data modality, including edges that point from diabetes and obesity to hypertension and hypercholesterolemia, from age to anxiety (which also had an edge toward depression), and from sex to diabetes, smoking, anxiety, hypercholesterolemia, and breastfed. In addition, we observed connections among different SNP groups, including SNPs that map to the genes *TMEM175*, *NSF*, *LRRC37A2*, *SNCA* and *KANSL1*. We also observed connections between genetic and non-genetic variables, namely an edge from SNP *rs1372519* located in the *SNCA* locus towards BMI and from *rs2230288*, located in the *GBA* locus, and *rs28399664*, located in the *BCAM* locus, towards exposure to air pollution.

Mendelian randomization identifies causal impact of hypertension on PDD

After calibrating the effects of the SNPs from the summary statistics and applying our instrument selection thresholds, only daytime sleep, diabetes, obesity, smoking, hearing loss, and hyperten- sion had SNPs that met the criteria: 37 for daytime sleep, 293 for diabetes, 3 for obesity, 22 for smoking status, 8 for hearing loss, and 463 for hypertension, which makes them suitable for MR analysis. IVW analysis indicated that hypertension increased the risk of developing Parkinson's disease (PD) dementia, with a causal estimate of 0.223 [se = 0.06, $p = 0.0005$], and diabetes also presented a risk with a causal estimate of 0.138 [se = 0.03, $p = 0.0004$]. Daytime sleep did not show a significant risk of dementia due to PD [$p = 0.3$], nor did obesity [$p = 0.87$], hearing loss [$p = 0.35$] or smoking status [$p = 0.27$]. The detailed results of our MR analysis are found in Supplementary Tables 4 and 5.

MR-Egger sensitivity analysis confirmed a significant causal effect of hypertension on PDD with an estimate of 0.3 [se = 0.19, $p = 0.04$] but did not reach significance for diabetes [$p = 0.22$], daytime sleep [$p = 0.27$], obesity [$p = 0.64$] or smoking status [$p = 0.78$]. This may be due to MR-Egger's lower power compared to IVW analysis due to the inclusion of the intercept term. No comorbidities showed signs of directional pleiotropy, as indicated by insignificant p-values for the MR-Egger intercept. MR-PRESSO analysis indicated horizontal pleiotropy for diabetes [rss = 338.7, $p = 0.03$], but causal effects remained significant after correction for outliers, with an estimate of 0.18 [se = 0.03, $p = 0.000002$] and no outlier distortion [$p = 0.3$]. The global test did not show signs of horizontal pleiotropy for hypertension or daytime sleep. Together, our MR analysis confirmed a potential causal link between hypertension and the risk of dementia from PD. Furthermore, our MR analysis indicates a likely causal impact of type 2 diabetes on the risk of developing PDD.

Discussion

To our knowledge, our study is the first to explore the predictability of PDD in PwPD and the interplay of various genetic and non-genetic factors in UKB. We developed a machine learning model that predicts PDD in people with Parkinson's disease (PwPD), achieving an average performance of 0.62 AUC. Although this performance may not be suitable for clinical use, it could be valuable for patient stratification to reduce sample sizes in clinical trials[30]. Additionally, it allowed us to evaluate the contributions of different data modalities, demonstrating that genetic factors had the most significant cumulative impact on the risk of PDD, followed by demographics and comorbidities.

The primary predictor of the model was the polygenic risk score PGS4281, calculated for all-cause dementia using 110 risk variants[31]. In our study, this score alone allowed us to predict PDD with an AUC of 57.5%.

Interestingly, we found SNPs in a gene cluster associated with AD risk to be predictive of PDD, namely *rs769449*, which is located in the Apolipoprotein E *APOE* locus. Although the APOE $\epsilon 4$ allele is well recognized as a risk factor for sporadic Alzheimer's disease (AD)[32], recent research indicates

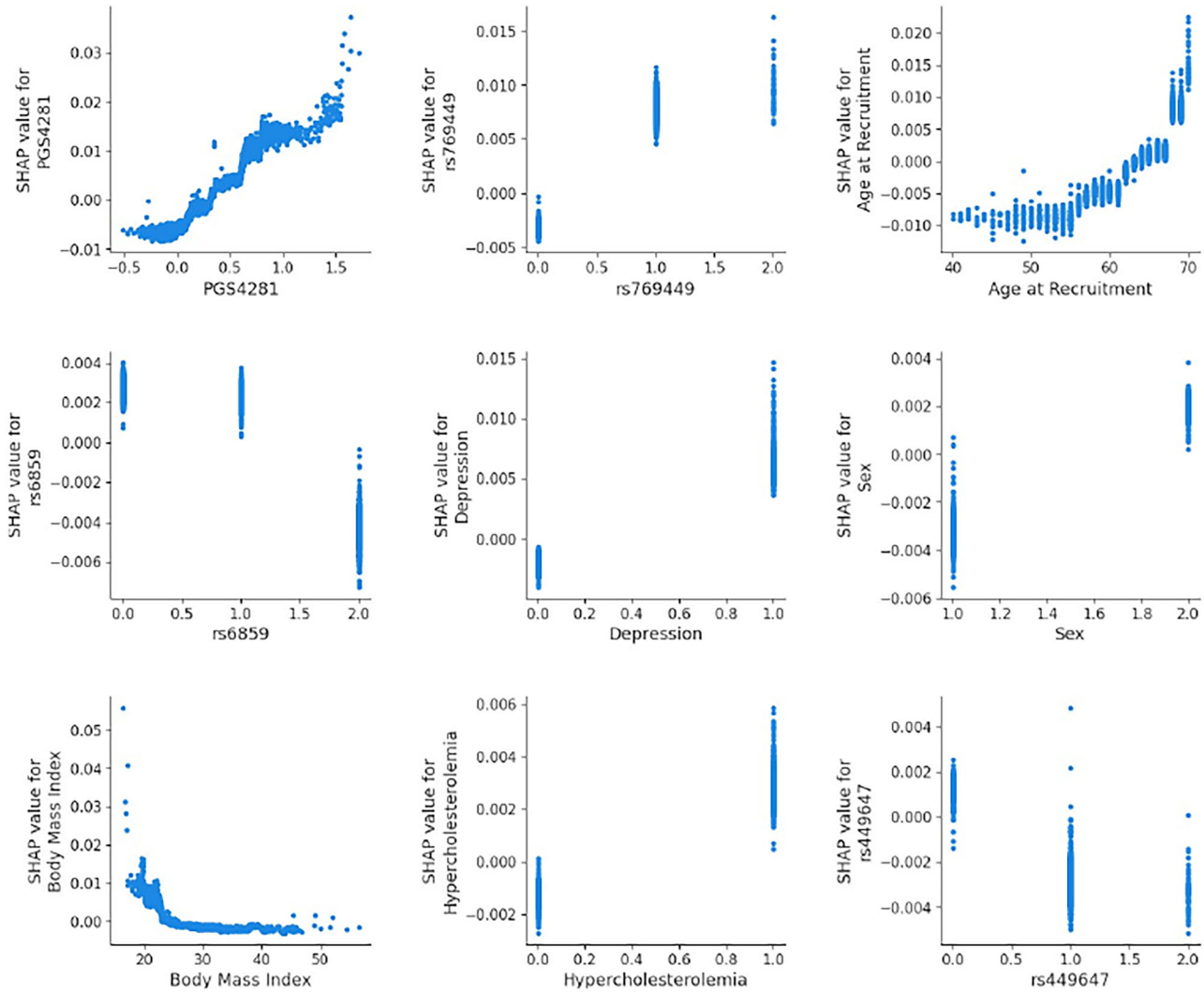

Fig. 5 | SHAP dependence plots. SHAP dependence plots for the top contributing features in the model. Each subplot displays the SHAP value (impact on model output) on the vertical axis against the corresponding feature value on the horizontal axis. These plots illustrate both the direction and magnitude of each feature's contribution to individual predictions. Non-linear relationships and feature interactions are observable in continuous variables such as PGS4281, Age at Recruitment, and Body Mass Index.

that it may also play a role in Parkinson's disease (PD)-related neurodegeneration by influencing α-synuclein pathology and its associated toxicity[33]. While α-synuclein aggregation is a hallmark of PD, a subset of PwPDD also exhibit AD-type pathology including amyloid-β plaques and tau neurofibrillary tangles[34,35]. As AD-type co-pathology is common in PDD, this aspect might point to the existence of subtypes of PDD and suggests further research to disentangle this heterogeneity.

As expected, age was among the main predictors in this model, as dementia is generally more prevalent in older people, and PwPD can develop dementia years after motor symptoms appear. However, age alone allowed one only to predict PDD slightly above the change level (AUC 56%), highlighting the need to consider different risk factors in combination.

The variant *rs6859* which is located in the *NECTIN2* locus was among the top predictors and has previously been linked to deterioration of cognitive abilities in adults[36]. We observed a strong influence of sex as a predictor, which is consistent with reported statistics indicating that men are twice as likely as women to develop PDD[37].

Another relevant predictor was BMI, which had an inverse relation to the probability for PDD. Higher BMI has previously been linked to a protective effect against fast cognitive decline and dementia development in PwPD[38–40]. However, this association may be partially confounded by disease-related complications in later stages of PD, where unintentional weight loss is common due to metabolic changes, dysphagia, and disease progression. This pattern is evident in the SHAP analysis, where age and BMI demonstrate opposing effect directions.

We identified various comorbidities as significant predictors, including depression, anxiety, hypertension, hypercholesterolemia, and excessive daytime sleepiness. Depression and anxiety are recognized non-motor symptoms of Parkinson's disease (PD) that can precede motor symptoms[41]. Excess daytime sleepiness is more prevalent in patients with PD with dementia compared to those with normal cognitive performance[42]. In addition, both hypertension and hypercholesterolemia are associated with an increased risk of dementia[43,44].

Looking at important genetic factors, we observed variants located in the *APOE* locus, including *rs449647* and *rs7412*. Furthermore, the variant *rs157580*, located in the *TOMM40* locus, and *rs60049679*, located in the *APOC1* locus, which make up the well-known Alzheimer's disease risk gene cluster, *APOE-TOMM40-APOC1*[45].

A notable predictor was tea intake. A previous MR analysis indicated that green tea consumption may slow PD progression and protect against dementia[46]. In contrast, another MR analysis found that the consumption of more than 13 cups of tea per day was associated with an increased risk of AD

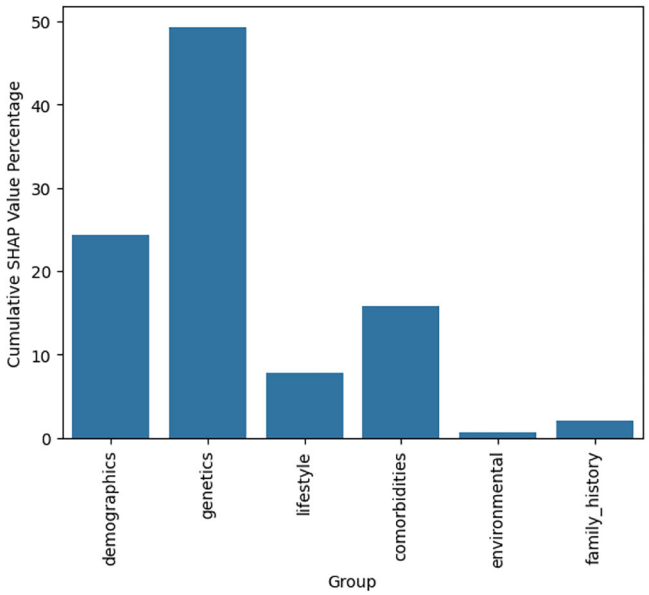

Fig. 6 | SHAP contribution by feature group. Cumulative SHAP value percentages grouped by data modality. The bar plot summarizes the relative contribution of each feature group to the model's output, with SHAP values aggregated within each category and normalized to sum to 100%. This grouping facilitates comparison of the influence of broader data domains.

dementia[47], possibly due to pesticide residues[48]. Similarly, water intake was among the predictors and could be attributed to lower water consumption in PwPD due to dysphagia[49]. Dehydration has been shown to accelerate cognitive decline in other studies[50–52], which could explain the effects of water intake as a variable in our analysis.

We examined interactions between phenotype-related variables and genotype-related features using BN structure learning. The expected interactions included the effect of age on hypertension, daily physical activity, and daytime sleepiness. Unexpected connections were also found, such as between maternal smoking and childhood obesity, possibly reflecting social deprivation. Furthermore, we observed links between sex, smoking status, and anxiety, as the prevalence of smoking and anxiety disorders varies between men and women[53,54]. Some interactions involved variables from different data modalities. For example, the variant *rs1372519* in the *SNCA* locus was associated with body mass index. Although individuals with Parkinson's disease (PwPD) generally have a lower BMI[55], a direct connection between the *SNCA* mutation and the differences in BMI remains undetermined. Moreover, two edges from *rs2230288* in the *GBA* locus and *rs28399664* in the *BCAM* locus indicated exposure to air pollution. The susceptibility of PwPD to certain gene mutations in regard to air pollution is not well documented, although similar gene-environment interactions have been identified in other diseases, such as chronic obstructive pulmonary disease (COPD)[56].

Lastly, we used MR analysis to examine the causal effects of type 2 diabetes, hypertension, hypercholesterolemia, obesity, and daytime sleepiness on the risk of developing PDD. MR analysis for other comorbidities was not possible, as none of the SNPs obtained from their GWAS summary statistics passed our instrument selection criteria. Our analysis revealed a putative causal effect of hypertension and type 2 diabetes. Studies have shown that elevated blood pressure in middle-aged populations (40–65 years) may be associated with a higher risk of cognitive impairment, while lower blood pressure in an elderly population (≥65) has been associated with a higher prevalence of dementia[57]. Another study has shown that hypertension is related to severe basal ganglia dilated perivascular space (BGdPVS), which in turn is associated with white matter hyperintensities (WMH) and cognitive decline in Parkinson's disease (PD)[58]. Similarly, previous studies show that type 2 diabetes is associated with faster motor and cognitive impairment in PD[59], and that long-term variability in blood glucose increases the risk of dementia in PD patients[60].

Our study has several limitations, including the reliance on ICD-coded reports of PD and PDD in UKB data, which may introduce unknown errors, particularly in distinguishing PDD from other types of dementia, such as Lewy Body Disorder. Another key limitation is the use of genotyping arrays rather than whole-genome or whole-exome sequencing, which may omit important genetic factors such as rare GBA variants. Furthermore, the generalizability of our findings is affected by the fact that most of the participants were white British, older on average, and predominantly male. Furthermore, the sample size for PD and PDD in UKB is still relatively small compared to the general disease population. We were able to partially address this aspect by investigating the replicability of our machine learning model and the SHAP analysis using data from the PPMI study. However, the number of PDD patients with comorbidities investigated in the UKB was inadequate to replicate our MR analysis.

In conclusion, to our knowledge, our study is the first to explore the predictability of PDD in PwPD based on data from the UKB population study. We demonstrated that results PPMI with newly diagnosed PwPD were comparable, and the impact of individual features on model predictions was similar in both PPMI and UKB. Although prediction performance may overall be insufficient for clinical application, our innovative approach combining machine learning, BN structure learning, and MR analysis has shed light on the complex interactions among genetic predisposition, comorbidities, environment, and lifestyle factors contributing to PDD. In particular, we identified potential causal links between hypertension, diabetes, and PDD, suggesting that the management of blood pressure and glucose levels in the blood in individuals with Parkinson's might be suitable targets for preventive strategies.

Methods

Data sources

The UKB is a large-scale database and research resource, playing an important role in studying a wide spectrum of complex diseases and advancing their diagnosis, prevention, and treatment[61]. The cohort was released for research in April 2012, and currently includes cross-sectional data for more than 500,000 individuals from the UK. The initial assessments involved the collection of data on demographics, clinical history, and lifestyle. The participating individuals were also genotyped, providing valuable genetic data that can facilitate the study of the genetic architecture of complex diseases. The UK Biobank study received ethical approval from the National Information Governance Board for Health and Social Care (England and Wales), the Community Health Index Advisory Group (Scotland), and the North West Multi-centre Research Ethics Committee. Written informed consent was obtained from all participants.

To further validate our findings from UKB, we used data from the PPMI study, which provides multiple data modalities, including demographic and genetic components[62]. The data set was initially released in 2010, and currently includes data for more than 1500 patients. The collection of human data in the PPMI study adhered to all relevant ethical guidelines. The study was approved by the Institutional Review Board or Independent Ethics Committee at each participating site. In Europe, these included Attikon University Hospital (Greece), Hospital Clinic de Barcelona and Hospital Universitario Donostia (Spain), Innsbruck University (Austria), Paracelsus-Elena-Klinik Kassel/University of Marburg (Germany), Imperial College London (UK), Pitie´-Salpe^trie're Hospital (France), and the University of Salerno (Italy). In the United States, participating sites included Emory University, Johns Hopkins University, University of Alabama at Birmingham, PD and Movement Disorders Center of Boca Raton, Boston University, Northwestern University, University of Cincinnati, Cleveland Clinic Foundation, Baylor College of Medicine, Institute for Neurodegenerative Disorders, Columbia University Medical Center, Beth Israel Medical Center, University of Pennsylvania, Oregon Health and Science University, University of Rochester, University of California San Diego, and University of California San Francisco.

Patients selection and variables

We selected all UKB patients diagnosed with PD in the hospital according to the ICD-10 code G20 and those with Parkinson's Dementia (F02.3). This resulted in 3541 PwPD and 487 patients with PDD. Data were cross-sectional and included past diagnoses, comorbidities, as well as genetic, environmental, and lifestyle factors associated with the risk of PD, as identified in a large meta-analysis[63]. Following this publication, we constructed variables that reflect known risk factors associated with dementia[64], including stroke, type 2 diabetes, hypertension, depression, anxiety, hearing loss, visual loss, hypercholesterolemia and obesity. We also included daytime sleepiness/dozing which was strongly associated with incident PD diagnosis in another study[65], using measurements from the UKB field 1220 (daytime dozing/sleeping). All variables were checked for missing or ambiguous values and encoded properly to account for categorical or ordinal variables using one-hot encoding or label encoding. All variables are detailed in Supplementary Table 1.

For PPMI, we included PwPD and labelled them by their cognitive status to include individuals with mild cognitive impairment (MCI) or dementia as recorded in the PPMI cognitive state table (where code 2 indicates MCI and code 3 indicates dementia). Together, the data set included a total of 637 PwPD, among which 122 had MCI or dementia.

Use of genotype data

Genotyping was performed on all UKB participants using UKBiLEVE and UK Biobank Axiom arrays. The UKBiLEVE array was initially designed and ran on approximately 50,000 participants, which was followed by the UK Biobank Axiom array designed to run on the remaining 450,000 participants. Both arrays share over 95% common content. For PPMI, more than 1500 participants were genotyped using the Illumina NeuroX array, which combines approximately 240,000 exome variants and approximately 24,000 custom variants with a focus on neurological diseases.

Since PDD is defined as dementia due to PD, we hypothesized that certain genetic variants associated with PD could also contribute to the risk of PDD. To construct variables for our machine learning models, we thus obtained 3189 single-nucleotide polymorphisms (SNPs) associated with PD from the ieu-b-7 dataset[66]. In addition, variants associated with all-cause dementia might play a role. Hence, we obtained 924 dementia-associated SNPs from the finn-b-F5 Dementia[67] dataset available on the IEUGWAS database (8 April 2024)[68]. The obtained SNPs were not initially clumped, meaning they included many correlated variants that were not necessarily independent. After obtaining this initial set of SNPs, we performed SNP clumping to retain only the most disease-associated and relatively independent variants. We applied thresholds of $p < 5E-8$, $r^2 < 0.1$, and a physical distance of 250 kb to account for potential differences in the linkage disequilibrium (LD) structure of our dataset. This process resulted in 10 highly associated SNPs for PD and 13 for dementia. In addition, we calculated two polygenic risk scores that are published in the Polygenic Risk Score Catalog (PGS)[69], namely PGS4281 for dementia[31] and PGS777 for PDD[70] using PLINK 2.0 to apply the effect sizes obtained from the PGS catalog to the genotype data obtained from ukbiobank while correctly handling the genotype dosages. Finally, we included 3 SNPs that have been associated with PDD from the DisGeNet gene-disease association network[71].

Machine learning models predicting PD dementia

We built three classification models to predict PDD within PwPD. This included gradient boosting[72], random forests[73] and elastic net penalized logistic regression[74]. XGBoost and random forests were chosen as classification models, because they are ensembles of decision trees, which are generally well-suited for multimodal tabular data, in which individual variables can follow different statistical distributions. Elastic net penalized logistic regression was added as a ground truth comparison due to its simplicity and ease of interpretation. The aim was to predict PDD using multimodal data, including SNP, polygenic risk scores, demographic characteristics, comorbidities, family history, and environmental and lifestyle-associated factors. The models were trained within a 5-fold nested cross-validation approach which splits data into a training set (80%) and a testing set (20%) while optimizing the model's hyperparameters within an inner cross-validation on the training set. Details on hyper-parameter optimization are presented in Supplementary Table 2. We evaluated the models' predictive performance by calculating the area under the receiver operating characteristic curve (AUC) across 20 repetitions of 5-fold nested cross-validation. To address label imbalance, class weights were applied during model training.

Making model predictions explainable

We calculated Shapley Additive Explanations (SHAP values) from the best performing machine learning model to understand the importance of features and the cumulative influence of different data modalities on predictions[75].

To better understand the interactions between all variables, we subsequently learned a Bayesian network (BN) from the data[76]. BNs are probabilistic graphical models that represent variables as nodes and conditional probabilistic dependencies between them as edges. We trained a BN on data from all subjects using non-parametric bootstrapping, which randomly selects samples 1000 times and learns a complete BN graph structure within each bootstrap using the tabu metaheuristic search method, which we have previously found to be the most accurate compared to multiple other BN structure learning methods for multimodal clinical data[77]. Continuous variables were discretized through clustering of k-means before learning the BN structure to account for the non-Gaussian nature of multiple features. BN structure learning was implemented using R-package bnlearn[78]. As each bootstrap sample resulted in a slightly different BN structure, a consensus structure was subsequently created by applying a conservative bootstrap probability threshold of 0.5, meaning that the edges of the averaged graph were found in at least 50% of all BN. This aims to improve the robustness and reliability of the estimates. Further details of BN learning can be found in Supplementary Table 3.

Mendelian randomization to estimate causal effects

MR assesses the causal relationship between an exposure (here: a comorbidity) and an outcome (here: PDD) using genetic variants as instrumental variables (IV). The variants used should satisfy three assumptions to be considered as an IV:

- The variant is associated with the exposure
- The variant is independent of confounding factors that confound the association of the exposure to the outcome
- The variant is independent of the outcome given the exposure and the confounding factors

In this regard, the genetic variants associated with exposure can be used as proxies to explain how exposure can influence the outcome of interest. For our study, we used MR to assess the putative causal effect of comorbidities on the risk of PDD.

We obtained genetic variants for our exposures using the IEU-OpenGWAS API (ieugwasr package[79]) from summary datasets for stroke (ebi-a-GCST005838)[80], type-2 diabetes (ebi- a-GCST90018926)[81], hypertension (ebi-a-GCST90038604)[82], depression (ukb-b-12064), daytime sleeping/dozing (ukb-b-5776)[83], anxiety (ukb-a-82)[84], hypercholesterolemia (finn-b-E4 HYPERCHOL), hearing loss (finn-b-H8 CONSENHEARINGLOSS), visual loss (finn-b-H7 VISUALDISTBLIND), obesity (finn-b-E4 OBESITY)[67], and smoking status (ebi-a-GCST90029014)[85]. The selection of summary datasets was based on the availability of top hits. We conducted one-sample MR by assessing the associations between genetic variants and both the exposures and the outcome at the individual level. We calibrated SNP exposure estimates to our phenotypes and generated SNP outcome estimates through association testing, adjusting for age and sex. The instruments were selected by clumping to ensure that the SNPs were strongly associated with exposures and independent of each other, applying thresholds of $p < 5E-8$, $r^2 < 0.1$ for linkage disequilibrium and a

genomic window of 250 kb. We estimated causal effects using the inverse variance weighted method (IVW) with fixed effects, which calculates the causal effect of exposure X on outcome Y. The causal effect ratio using a genetic variant i is X_i/Y_i, with the standard error estimated by the delta method as σ_{Yi}/X_i.

The IVW estimate combines ratio estimates of the variants in a fixed-effect meta-analysis model.

$$\widehat{\beta_{\mathrm{IVW}}} = \frac{\sum_k X_k Y_k \sigma_{Y_k}^{-2}}{\sum_k X_k^2 \sigma_{Y_k}^{-2}} \quad (1)$$

The success of MR depends on the three previously mentioned assumptions, which require a sensitivity analysis for the evaluation of the results[86]. We utilized MR-Egger and MR-PRESSO for this analysis. The MR-Egger method evaluates directional pleiotropy, assessing whether genetic factors have average pleiotropic effects on the outcome that differ from zero. Pleiotropy in human genetics can manifest in various forms, such as a single variant influencing multiple traits or a causal locus affecting a trait through another one. Horizontal pleiotropy occurs when a variant directly or indirectly impacts the target outcome, often by influencing other traits that causally affect it. In this context, MR-Egger provides a reliable estimate of the causal effect under the assumption of InSIDE (Instrument Strength Independent of Direct Effect)[87]. In this work, MR-PRESSO was used to investigate horizontal pleiotropy in multi-instrument summary level MR[88], which involves three steps: a global test for horizontal pleiotropy detection, removal of outliers for correction, and a significance test before and after removal of outliers.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

This research has been conducted using the UK Biobank resource under application number 67829. The genetic and phenotype datasets are not publicly available but can be accessed via the UK Biobank data access process. More details are available at http://www.ukbiobank.ac.uk/register-apply/. PPMI data are publicly available from the Parkinson's Progression Markers Initiative (PPMI) database www.ppmi-info.org/access-data-specimens/download-data, RRID:SCR 006431. For up-to-date information on the study, visit www.ppmi-info.org.

Code availability

The underlying code used for model training, Mendelian randomization and learning the Bayesian networks in this study will be published on GitHub upon acceptance of the paper under the CC BY-NC-ND 4.0 license and can be accessed via https://github.com/mosala777/Predicting-Dementia-in-PwPD.

Received: 2 January 2025; Accepted: 28 April 2025;
Published online: 13 May 2025

References

1. Feigin, V. L. et al. Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Neurol.* **16**, 877–897 (2017).
2. Blauwendraat, C., Nalls, M. A. & Singleton, A. B. The genetic architecture of Parkinson's disease. *Lancet Neurol*. **19**, 170–178 (2020).
3. Tsalenchuk, M., Gentleman, S. M. & Marzi, S. J. Linking environmental risk factors with epigenetic mechanisms in Parkinson's disease. *NPJ Parkinsons Dis.* **9**, 123 (2023).
4. McNaught, K. S. et al. Effects of isoquinoline derivatives structurally related to 1-methyl- 4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) on mitochondrial respiration. *Biochem. Pharmacol.* **51**, 1503–1511 (1996).
5. Pouchieu, C. et al. Pesticide use in agriculture and Parkinson's disease in the AGRICAN cohort study. *Int. J. Epidemiol.* **47**, 299–310 (2018).
6. Jo, S. et al. Association of NO 2 and Other Air Pollution Exposures With the Risk of Parkinson Disease. *JAMA Neurol.* **78**, 800 (2021).
7. Hughes, K. C. et al. Intake of dairy foods and risk of Parkinson disease. *Neurol* **89**, 46–52 (2017).
8. Belvisi, D. et al. Modifiable risk and protective factors in disease development, progression and clinical subtypes of Parkinson's disease: What do prospective studies suggest? *Neurobiol. Dis.* **134**, 104671 (2020).
9. Aarsland, D. et al. Mild cognitive impairment in Parkinson disease. *Neurol* **75**, 1062–1069 (2010).
10. Williams-Gray, C. H. et al. The CamPalGN study of Parkinson's disease: 10-year outlook in an incident population-based cohort. *J. Neurol. Neurosurg. Psychiatry* **84**, 1258–1264 (2013).
11. Anang, J. B. et al. Predictors of dementia in Parkinson disease. *Neurol* **83**, 1253–1260 (2014).
12. Amer, H. et al. Genetic Influences on Cognition in Idiopathic Parkinson's Disease. *Neurol. Res. Int.* **2018**, 5603571 (2018).
13. Planas-Ballve, A. & Vilas, D. Cognitive Impairment in Genetic Parkinson's Disease. *Parkinsons Dis*. **2021**, 8610285 (2021).
14. Ba¨ckstro¨ M., D. et al. Prediction and early biomarkers of cognitive decline in Parkinson disease and atypical parkinsonism: a population-based study. *Brain Commun* **4**, fcac040 (2022).
15. Svenningsson, P., Westman, E., Ballard, C. & Aarsland, D. Cognitive impairment in patients with Parkinson's disease: diagnosis, biomarkers, and treatment. *Lancet Neurol.* **11**, 697–707 (2012).
16. Kempster, P. A., O'Sullivan, S. S., Holton, J. L., Revesz, T. & Lees, A. J. Relationships between age and late progression of Parkinson's disease: a clinico-pathological study. *Brain* **133**, 1755–1762 (2010).
17. Ramezani, M. et al. Investigating the relationship between the SNCA gene and cognitive abilities in idiopathic Parkinson's disease using machine learning. *Sci. Rep.* **11**, 4917 (2021).
18. Kapasi, A. et al. A novel SNCA E83Q mutation in a case of dementia with Lewy bodies and atypical frontotemporal lobar degeneration. *Neuropathology* **40**, 620–626 (2020).
19. Szwedo, A. A. et al. ¡span Style="font-variant:small-caps;"¿ GBA ¡/span¿ and ¡span Style="font-variant:small-caps;"¿ APOE ¡/span¿ Impact Cognitive Decline in Parkinson's Disease: A 10-Year Population-Based Study. *Mov. Disord.* **37**, 1016–1027 (2022).
20. Huang, X., Chen, P., Kaufer, D. I., Tro¨ster, A. I. & Poole, C. Apolipoprotein E and Dementia in Parkinson Disease: A Meta-analysis. *Arch. Neurol.* **63**, 189 (2006).
21. Williams-Gray, C. H. et al. The distinct cognitive syndromes of Parkinson's disease: 5 year follow-up of the CamPalGN cohort. *Brain* **132**, 2958–2969 (2009).
22. Seto´-Salvia, N. et al. Dementia Risk in Parkinson Disease. *Arch. Neurol*. **68**, 359–364 (2011).
23. Aarsland, D. et al. Parkinson disease-associated cognitive impairment. *Nat. Rev. Dis. Prim.* **7**, 47 (2021).
24. Neumann, J. et al. Glucocerebrosidase mutations in clinical and pathologically proven Parkinson's disease. *Brain* **132**, 1783–1794 (2009).
25. Seto´-Salvia, N. et al. Glucocerebrosidase mutations confer a greater risk of dementia during Parkinson's disease course. *Mov. Disord.* **27**, 393–399 (2012).
26. Phongpreecha, T. et al. Multivariate prediction of dementia in Parkinson's disease. *NPJ Parkinsons Dis.* **6**, 20 (2020).
27. Harvey, J. et al. Machine learning-based prediction of cognitive outcomes in de novo Parkinson's disease. *NPJ Parkinsons Dis.* **8**, 150 (2022).

28. Wooten, G. F. Are men at greater risk for Parkinson's disease than women? *J. Neurol. Neurosurg. Psychiatry* **75**, 637–639 (2004).
29. Taylor, K. S. M., Cook, J. A. & Counsell, C. E. Heterogeneity in male to female risk for Parkinson's disease. *J. Neurol. Neurosurg Psychiatry* **78**, 905–906 (2007).
30. Ha¨hnel, T. et al. Progression subtypes in Parkinson's disease identified by a data-driven multicohort analysis. *NPJ Parkinsons Dis.* **10**, 95 (2024).
31. Ohta, R., Tanigawa, Y., Suzuki, Y., Kellis, M. & Morishita, S. A polygenic score method boosted by non-additive models. *Nat. Commun.* **15**, 4433 (2024).
32. Van der Kant, R., Goldstein, L. S. B. & Ossenkoppele, R. Amyloid-β-independent regulators of tau pathology in Alzheimer disease. *Nat. Rev. Neurosci.* **21**, 21–35 (2020).
33. Davis, A. A. et al. APOE genotype regulates pathology and disease progression in synucleinopathy. *Sci. Transl. Med.* **12**, eaay3069 (2020).
34. Toledo, J. B. et al. Pathological α-synuclein distribution in subjects with coincident Alzheimer's and Lewy body pathology. *Acta Neuropathologica* **131**, 393–409 (2016).
35. Irwin, D. J. et al. Neuropathological and genetic correlates of survival and dementia onset in synucleinopathies: a retrospective analysis. *Lancet Neurol.* **16**, 55–65 (2017).
36. Rajendrakumar, A. L., Arbeev, K. G., Bagley, O., Yashin, A. I. & Ukraintseva, S. The SNP rs6859 in NECTIN2 gene is associated with underlying heterogeneous trajectories of cognitive changes in older adults. *BMC Neurol.* **24**, 78 (2024).
37. Mouton, A. et al. Sex ratio in dementia with Lewy bodies balanced between Alzheimer's disease and Parkinson's disease dementia: a cross-sectional study. *Alzheimers Res. Ther.* **10**, 92 (2018).
38. Yoo, H. S., Chung, S. J., Lee, P. H., Sohn, Y. H. & Kang, S. Y. The Influence of Body Mass Index at Diagnosis on Cognitive Decline in Parkinson's Disease. *J. Clin. Neurol.* **15**, 517–526 (2019).
39. Rahmani, J. et al. Body mass index and risk of Parkinson, Alzheimer, Dementia, and Dementia mortality: a systematic review and dose–response meta-analysis of cohort studies among 5 million participants. *Nutr. Neurosci.* **25**, 423–431 (2022).
40. Natale, G., Zhang, Y., Hanes, D. W. & Clouston, S. A. Obesity in Late-Life as a Protective Factor Against Dementia and Dementia-Related Mortality. *Am. J. Alzheimers Dis. Other Demen* **38**, 15333175221111658 (2023).
41. Schrag, A. & Taddei, R. N. in *International Review of Neurobiology* 623–655 (Elsevier, 2017).
42. Liu, H. et al. Excessive Daytime Sleepiness in Parkinson's Disease. *Nat. Sci. Sleep* **14**, (2022).
43. McGrath, E. R. et al. Blood pressure from mid- to late life and risk of incident dementia. *Neurology* **89**, 2447–2454 (2017).
44. Iwagami, M. et al. Blood cholesterol and risk of dementia in more than 1·8 million people over two decades: a retrospective cohort study. *Lancet Healthy Longev.* **2**, e498–e506 (2021).
45. Kamboh, M. I. et al. Genome-wide association study of Alzheimer's disease. *Transl. Psychiatry* **2**, e117 (2012).
46. Li, C., Lin, J., Yang, T. & Shang, H. Green Tea Intake and Parkinson's Disease Progression: A Mendelian Randomization Study. *Front. Nutr.* **9**, 848223 (2022).
47. Sun, Y. et al. Extra cup of tea intake associated with increased risk of Alzheimer's disease: Genetic insights from Mendelian randomization. *Front. Nutr.* **10**, 1052281 (2023).
48. De Andrade Arruda Fernandes, I. et al. The bitter side of teas: Pesticide residues and their impact on human health. *Food Chem. Toxicol*. **179**, 113955 (2023).
49. Suttrup, I. & Warnecke, T. Dysphagia in Parkinson's Disease. *Dysphagia* **31**, 24–32 (2016).
50. Lauriola, M. et al. Neurocognitive Disorders and Dehydration in Older Patients: Clinical Experience Supports the Hydromolecular Hypothesis of Dementia. *Nutrients* **10**, 562 (2018).
51. Aslan Kirazoglu, D. et al. The relationship between dehydration and etiologic subtypes of major neurocognitive disorder in older patients. *Eur. Geriatr. Med.* **15**, 1159–1168 (2024).
52. Nishi, S. K. et al. Water intake, hydration status and 2-year changes in cognitive performance: a prospective cohort study. *BMC Med.* **21**, 82 (2023).
53. Waldron, I. Patterns and causes of gender differences in smoking. *Soc. Sci. Med.* **32**, 989–1005 (1991).
54. McLean, C. P., Asnaani, A., Litz, B. T. & Hofmann, S. G. Gender differences in anxiety disorders: Prevalence, course of illness, comorbidity and burden of illness. *J. Psychiatr. Res.* **45**, 1027–1035 (2011).
55. Van der Marck, M. A. et al. Body mass index in Parkinson's disease: A meta-analysis. *Parkinsonism Relat. Disord.* **18**, 263–267 (2012).
56. Melbourne, C. A. et al. Genome-wide gene-air pollution interaction analysis of lung function in 300,000 individuals. *Environ. Int* **159**, 107041 (2022).
57. Sierra, C. Hypertension and the Risk of Dementia. *Front. Cardiovasc. Med.* **7**, 5 (2020).
58. Shin, N.-Y. et al. Adverse effects of hypertension, supine hypertension, and perivascular space on cognition and motor function in PD. *NPJ Parkinsons Dis.* **7**, 69 (2021).
59. Chohan, H. et al. Type 2 Diabetes as a Determinant of Parkinson's Disease Risk and Progression. *Mov. Disord.* **36**, 1420–1429 (2021).
60. Kang, S. H. et al. Fasting glucose variability and risk of dementia in Parkinson's disease: a 9-year longitudinal follow-up study of a nationwide cohort. *Front. Aging Neurosci.* **15**, 1292524 (2024).
61. Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, e1001779 (2015).
62. Marek, K. et al. The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* **95**, 629–635 (2011).
63. Noyce, A. J. et al. Meta-analysis of early nonmotor features and risk factors for Parkinson disease. *Ann. Neurol*. **72**, 893–901 (2012).
64. Livingston, G. et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. *Lancet* **404**, 572–628 (2024).
65. Jacobs, B. M. et al. Parkinson's disease determinants, prediction and gene–environment interactions in the UK Biobank. *J. Neurol. Neurosurg. Psychiatry* **91**, 1046–1054 (2020).
66. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
67. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
68. Elsworth, B. et al. *The MRC IEU OpenGWAS data infrastructure. bioRxiv* https://research-information.bris.ac.uk/en/publications/the-mrc-ieu-opengwas-data-infrastructure (2020).
69. Lambert, S. A. et al. The Polygenic Score Catalog: new functionality and tools to enable FAIR research. *medRxiv*, https://www.medrxiv.org/content/10.1101/2024.05.29.24307783v1 (2024).
70. Liu, G. et al. Genome-wide survival study identifies a novel synaptic locus and polygenic score for cognitive progression in Parkinson's disease. *Nat. Genet* **53**, 787–793 (2021).
71. Pin˜ero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. **48**, D845–D855 (2020).
72. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 0090–5364 (2001).
73. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
74. Zou, H. & Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
75. Lundberg, S. M. & Lee, S.-I. *A unified approach to interpreting model predictions in Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2017), 4768–4777.

76. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning* (The MIT Press, 2009).
77. Sood, M. et al. Realistic simulation of virtual multi-scale, multi-modal patient trajectories using Bayesian networks and sparse auto-encoders. *Sci. Rep.* **10**, 10971 (2020).
78. Scutari, M. & Denis, J.-B. *Bayesian Networks with Examples in R* 2nd. (Chapman and Hall, 2021).
79. Hemani, G., Elsworth, B., Palmer, T. & Rasteiro, R. *ieugwasr: Interface to the 'OpenGWAS' Database API* R package version 1.0.1 https://mrcieu.github.io/ieugwasr/ (2024).
80. Malik, R. et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524–537 (2018).
81. Sakaue, S. et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
82. Do¨nertas, H. M., Fabian, D. K., Fuentealba, M., Partridge, L. & Thornton, J. M. Common genetic associations between age-related diseases. *Nat. Aging* **1**, 400–412 (2021).
83. Mitchell, R. et al. *MRC IEU UK Biobank GWAS pipeline version 2. University of Bristol* https://data.bris.ac.uk/data/dataset/pnoat8cxo0u52p6ynfaekeigi (2019).
84. Howrigan, D. P. et al. *Nealelab/UK Biobank GWAS: version 2. Neale Lab* http://www.nealelab.is/uk-biobank/ (2023).
85. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
86. Zhang, W. & Ghosh, D. A General Approach to Sensitivity Analysis for Mendelian Randomization. *Stat. Biosci.* **13**, 34–55 (2021).
87. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* **32**, 377–389 (2017).
88. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).

Acknowledgements

This research has been conducted using the UK Biobank resource under application 67829. Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database www.ppmi-info.org/access-dataspecimens/download-data, RRID:SCR_006431. For up-to-date information on the study, visit www.ppmi-info.org. 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/.

Author contributions

M.A.: data curation, formal analysis, investigation, methodology, visualizing, writing - original draft; T.H.: data curation, formal analysis, methodology, writing - review & editing; P.M.: conceptualization, writing - review & editing; J.K.: conceptualization, writing - review & editing; H.F.: conceptualization, methodology, supervision, writing - review & editing. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

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/s41531-025-00983-4.

Correspondence and requests for materials should be addressed to Mohamed Aborageh or Holger Fröhlich.