Gene–Environment Interactions for Parkinson's Disease

Alexandra Reynoso, MSc [1], Roberta Torricelli, BSc [2], Benjamin Meir Jacobs, MRCP, MSc [2], Jingchunzi Shi, PhD,[1] Stella Aslibekyan, PhD,[1] Lucy Norcliffe-Kaufmann, PhD,[1] Alastair J Noyce, MRCP, PhD [2] and Karl Heilbron, PhD [3,4]

Objective: Parkinson's disease (PD) is a neurodegenerative disorder with complex etiology. Multiple genetic and environmental factors have been associated with PD, but most PD risk remains unexplained. The aim of this study was to test for statistical interactions between PD-related genetic and environmental exposures in the 23andMe, Inc. research dataset.

Methods: Using a validated PD polygenic risk score and common PD-associated variants in the *GBA* gene, we explored interactions between genetic susceptibility factors and 7 lifestyle and environmental factors: body mass index (BMI), type 2 diabetes (T2D), tobacco use, caffeine consumption, pesticide exposure, head injury, and physical activity (PA).

Results: We observed that T2D, as well as higher BMI, caffeine consumption, and tobacco use, were associated with lower odds of PD, whereas head injury, pesticide exposure, *GBA* carrier status, and PD polygenic risk score were associated with higher odds. No significant association was observed between PA and PD. In interaction analyses, we found statistical evidence for an interaction between polygenic risk of PD and the following environmental/lifestyle factors: T2D ($p = 6.502 \times 10^{-8}$), PA ($p = 8.745 \times 10^{-5}$), BMI ($p = 4.314 \times 10^{-4}$), and tobacco use ($p = 2.236 \times 10^{-3}$). Although BMI and tobacco use were associated with lower odds of PD regardless of the extent of individual genetic liability, the direction of the relationship between odds of PD and T2D, as well as PD and PA, varied depending on polygenic risk score.

Interpretation: We provide preliminary evidence that associations between some environmental and lifestyle factors and PD may be modified by genotype.

ANN NEUROL 2024;00:1–11

Parkinson's disease (PD) may be the fastest growing neurological disorder worldwide, with prevalence of 1% to 2% in the age >60 years population.[1] It is characterized by chronic, progressive neuronal loss and intracellular alpha-synuclein inclusions (Lewy bodies). The genetic architecture of PD involves contributions from both common, modest effect-size alleles, and rarer, monogenic forms identified via linkage analyses of affected families, such as *SNCA*, *PINK1*, *PARK7*, *LRRK2*, and *PRKN*.[2] Over the past decade, genome-wide association studies (GWAS) have identified susceptibility loci involved in complex disease, broadening our understanding of the genetic basis of PD.[3,4]

Among the genes observed to play a role in PD etiology, the glucocerebrosidase (*GBA*) gene locus, also known for its role in Gaucher's disease, has emerged as a notable risk factor for sporadic disease. Approximately 10 to 15% of European PD patients carry a PD-associated

View this article online at wileyonlinelibrary.com. DOI: 10.1002/ana.26852

Received Jun 16, 2023, and in revised form Dec 6, 2023. Accepted for publication Dec 6, 2023.

Address correspondence to Alastair J. Noyce, Centre for Preventive Neurology, Wolfson Institute of Population Health, Faculty of Medicine and Dentistry, Queen Mary University of London, London, UK. E-mail: a.noyce@qmul.ac.uk; Karl Heilbron, Department of Psychiatry and Psychotherapy, Charité Universitätsmedizin, Berlin, Germany. E-mail: karl.heilbron@charite.de

Alexandra Reynoso and Roberta Torricelli contributed equally to this work. Alastair J. Noyce and Karl Heilbron contributed equally to this work.

From the [1]23andMe, Inc., Sunnyvale, CA, USA; [2]Center for Preventive Neurology, Wolfson Institute of Population Health, Faculty of Medicine and Dentistry, Queen Mary University of London, London, UK; [3]Department of Psychiatry and Psychotherapy, Charité Universitätsmedizin, Berlin, Germany; and [4]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA

GBA variant, with the presence of multiple, aggregated variants increasing PD risk by up to 15-fold more compared with the general population, depending on severity.[5,6] Pathogenic variants of the *GBA* gene are associated with a distinctive phenotype, including younger onset and more severe non-motor features.[2,7] Of the common risk variants in *GBA*, the most prevalent are N370S, E326K, and T369M, particularly in individuals with Ashkenazi Jewish descent.[7]

The cumulative impact of common risk alleles can be summarized at an individual level into a polygenic risk score (PRS), which has been successfully applied in research settings to identify putative interactions between genetic and lifestyle risk factors in the pathogenesis of PD. Calculation of PRS has enabled us to explain 16 to 36% of PD heritability.[4] This information on genetic risk can then be further characterized in conjunction with modifiable phenotypic risk factors identified through epidemiological studies, thereby advancing our understanding of complex disease etiology.[8]

Indeed, it is believed that lifestyle and environmental exposures contribute to the incidence of sporadic disease, with some playing a greater role than others. There is significant evidence implicating head injuries, pesticides exposure, tobacco, and caffeine consumption in idiopathic PD.[9–11] Similarly, conditions, such as metabolic syndrome, as well as type 2 diabetes (T2D) and BMI individually, have also been described to influence disease development in several studies.[12,13]

Previous studies have examined interactions between *LRRK2* and various lifestyle factors, such as tobacco, black tea consumption, and NSAID use, all of which were shown to be associated with a decreased risk of disease penetrance.[14,15] Other studies have examined evidence for interactions between a PD PRS and common health risk factors, such as diabetes, alcohol, and tobacco consumption, which revealed a highly complex genetic etiology with variable gene-by-environment interactions.[16]

In this large population-based cross-sectional study, we explored the interaction between genetic risk factors for PD – both in terms of polygenic risk and pathogenic *GBA* variants – and various lifestyle and environmental factors.

Methods

Participants

The dataset consisted of customers of 23andMe, Inc., a direct-to-consumer genetics company. Informed written consent to participate in research was obtained from all participants.

IRB Statement

Participants provided informed consent and volunteered to participate in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent (E&I) Review Services. As of 2022, E&I Review Services is part of Salus IRB (https://www.versiticlinicaltrials.org/salusirb).

Parkinson's Disease Study Cohort

23andMe's PD cohort was assembled from the customer database back in 2009 with continuous enrollment since. The project was designed to collect survey data in a cohort of consented participants who self-reported a diagnosis of PD.

Previous data from UK Biobank have suggested that self-reported PD is a reliable proxy for a PD diagnosis, showing a strong correlation with diagnostic codes for PD recorded in the Electronic Healthcare Records.[16] In a small internal validation cohort (n = 50), there was 100% concordance between self-reported PD in 23andMe and clinical assessment by a neurologist.[17]

The cohort has since expanded with additional recruitment collaborations with the Michael J. Fox Foundation and other PD patient advocacy groups.[18] Since its inception, participants who self-reported a diagnosis of PD diagnosis were targeted to complete a dedicated online PD survey that included a comprehensive series of questions designed to assess signs/symptoms of PD, risk factors, lifestyle habits, and lifetime environmental exposures. Controls were recruited from the pool of research-consented participants who did not report a diagnosis of PD or parkinsonism at entry or on follow-up surveys. PD cases and controls were removed from the analysis cohort if they reported: (1) a diagnosis of atypical parkinsonism (e.g. multiple system atrophy); (2) a history of severe vascular disease (stroke, deep vein thrombosis, or pulmonary embolism); or (3) a change in their diagnosis on subsequent survey.

Lifestyle and Environmental Factors

We prioritized 7 variables that have been repeatedly reported in previous PD epidemiological studies as modifiable environmental exposures or comorbidities: T2D, tobacco use, caffeine consumption, BMI, pesticide exposure, head injury, and physical activity (PA).[11,19] Answers from the self-reported survey questions were matched to each variable, and the data extracted. T2D was recorded as the presence of a previous diagnosis (yes/no). BMI was calculated from mass (kg) divided by height squared (m^2) and the quantile was normalized separately in men and women. Tobacco consumption was dichotomized by a smoking history of at least 100 cigarettes in a lifetime

(yes/no). Caffeine consumption was measured as daily milligrams of caffeine from any of the following: coffee, tea, soda, or energy drinks. Caffeine consumption was then transformed by log10(x + 75) to create a more normally distributed variable. Pesticide exposure was defined as use of pesticides, herbicides, fungicides, insecticides, rodenticides, or fumigants in a home or garden in a typical month (yes/no). Head injury history was recorded as having ever had a head injury or concussion as the result of a sporting activity, fall, violence, car accident, or other accidents that happened during childhood and adulthood (yes/no). Physical activity was measured as the amount of times per week a participant engaged in physical activity for >30 minutes. All variables were recorded cross-sectionally and assessed at the time of survey, which for most participants was after they had self-reported their diagnosis of PD. As such, the analysis is limited to exploring interactions in cross-sectional data rather than examining temporal associations.

Genotyping

DNA extraction and genotyping were performed on saliva samples by Clinical Laboratory Improvement Amendments-certified and College of American Pathologists-accredited clinical laboratories of Laboratory Corporation of America. Samples were genotyped on 1 of 5 genotyping platforms. The V1 and V2 platforms were variants of the Illumina HumanHap550+ BeadChip, and contained a total of ~560,000 single-nucleotide polymorphisms (SNPs), including ~25,000 custom SNPs selected by 23andMe. The V3 platform was based on the Illumina OmniExpress + BeadChip, and contained a total of ~950,000 SNPs and custom content to improve the overlap with our V2 array. The V4 platform is a fully custom array, and includes a lower redundancy subset of V2 and V3 SNPs with additional coverage of lower-frequency coding variation, and ~570,000 SNPs. The V5 platform is an Illumina Infinium Global Screening Array of ~640,000 SNPs supplemented with ~50,000 SNPs of custom content. Samples had minimum call rates of 98.5%.

SNPs were removed if they: (1) were only genotyped on the V1 and/or V2 platforms, (2) were located on chrM or chrY, (3) failed a test for parent–offspring transmission, (4) had a Hardy–Weinberg equilibrium $p < 10^{-20}$, (5) had a call rate <90%, (6) had $p < 10^{-50}$ in an ANOVA of genotype versus 20 equally-sized genotype date bins, (7) had $R^2 > 0.1$ in an ANOVA of genotype versus sex, or (8) had probes matching multiple genomic positions in the reference genome.

Imputation

We imputed variants using the Human Reference Consortium imputation reference panel[12] (32,488 samples, 39,235,157 SNPs). Multiallelic sites with N alternate alleles were split into N separate biallelic sites. We then removed any site whose minor allele appeared in only 1 sample. In preparation for imputation, we split each chromosome of the reference panel into chunks of no more than 300,000 variants, with overlaps of 10,000 variants on each side. We used a single batch of 10,000 individuals to estimate Minimac3 imputation model parameters for each chunk.

To generate phased participant data for the V1 to V4 platforms, we used an internally developed tool, Finch, which implements the Beagle graph-based haplotype phasing algorithm, modified to separate the haplotype graph construction and phasing steps.[20] Finch extends the Beagle model to accommodate genotyping error and recombination, to handle cases where there are no consistent paths through the haplotype graph for the individual being phased. We constructed haplotype graphs for all participants from a representative sample of genotyped individuals, and then performed out-of-sample phasing of all genotyped individuals against the appropriate graph.

SNPs with imputation $R^2 < 0.3$ were removed. SNP dosages were tested for platform batch effects between the V4 and V5 platforms via ANOVA and SNPs with $p < 10^{-50}$ were removed.

GBA Variants

We aggregated 3 known *GBA* PD risk-associated variants E326K (rs2230288, imputed, $r^2 = 0.994$), T369M (rs75548401, imputed, $r^2 = 0.747$), and N370S (rs76763715, genotyped) into a single binary variable (referred to hereafter as "*GBA* carrier status") denoting whether an individual was a carrier for any variant. Following the methods of previous groups,[21] variants E326K, T369M, and N370S were combined into a single *GBA* variable due to their similar effects on GCase activity in humans, reducing it by 18 to 46% on average.[22]

PRS Calculation

To select SNPs for inclusion into the PRS, we used summary statistics from the most recently published European-ancestry PD GWAS excluding 23andMe participants.[4] We restricted the analysis to common (minor allele frequency >1%), biallelic, autosomal, non-palindromic SNPs. SNPs were selected using the "clumping-and-thresholding (C + T)" approach. We used all possible combinations of 5 clumping thresholds

and 11 p value thresholds to generate 55 PRSs. SNPs were LD-clumped using PLINK (version 1.9) with a clumping distance of 250 kb and 5 clumping R^2 thresholds: 0.1, 0.2, 0.4, 0.6, and 0.8. The 503 European-ancestry samples from the 1,000 Genomes project were used as the LD reference. Using p values for each SNP's association with PD, we filtered SNPs using 11 p value thresholds: 5×10^{-8}, 5×10^{-5}, 5×10^{-4}, 0.005, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, and 1. Effect size estimates for the association of each variant with PD were obtained from the GWAS beta coefficient, representing the per-allele log odds ratio for PD.

We constructed 55 PRSs for 23andMe research participants as the weighted sum of risk allele counts for the SNPs selected in each of the 55 C + T profiles. We matched SNPs from the PD GWAS to 23andMe SNPs using the CPRA (chromosome, position, reference allele, alternative allele) format and excluding unmatched variants.[4] We harmonized SNP effect size estimates following the variant harmonization schema in Hartwig et al. 2016.[23] To ensure good SNP quality, we removed SNPs with imputation $R^2 < 0.5$ or a difference in minor allele frequency >30% between the PD GWAS variant and the 23andMe variant. For each of the 55 C + T profiles, PRS values were calculated for each individual, i, as $PRS_i = G_i^T \beta$, where β is an $N \times 1$ vector of harmonized weights for the N selected SNPs, and G_i is the corresponding $N \times 1$ vector of imputed dosages of the N selected SNPs in individual i. Finally, we standardized the PRS across all participants to have mean 0 and standard deviation 1.

To identify the best-performing PRS, we split our dataset into a 30% training test set and a 70% held-out test set. In the test set, we performed 55 logistic regressions using PD as the dependent variable, 1 of the 55 PRSs as the independent variable, and the following covariates: age (determined at the time the survey information was collected), sex, household income (inferred from zip code), and 5 principal components of genetic ancestry. We selected the PRS with the largest McFadden R^2 value for use in the 70% held-out set. The best-performing PRS used a clumping R^2 of 0.8 and a p-value threshold of 0.4. The corresponding C + T profile contained 603,976 SNPs before harmonization and 593,886 SNPs (98.3%) after harmonization.

Age and Sex-Matched Datasets

For each of the 7 lifestyle and environmental factors, we constructed age and sex-matched PD case–control datasets; including all individuals from the 70% held-out set with available data for the variable of interest. Specifically, we divided PD cases into 20 evenly populated age bins; determined the number of controls that fell into each corresponding age bin; divided the number of controls in a bin by the number of corresponding cases; determined the minimum number of available controls per case across all bins; and randomly selected this minimum number of age and sex-matched controls for each case. We also created a matched "full dataset" containing all PD cases regardless of whether data were available for the environment and lifestyle factors. We used the full dataset for all analyses that did not use data from any of the 7 environment and lifestyle factors. These case–control datasets only contained individuals with predominantly European ancestry, as the PRS was derived from a European ancestry GWAS.[24] Table 1 provides information on the sample size for each case–control dataset.

Statistical Analysis

Logistic regression models were run in our age and sex-matched datasets using PD status as the outcome, 1 of the environmental/lifestyle variables, and the following covariates: age (determined at the time the survey information was collected), sex, and 5 principal components of genetic ancestry. Unlike the PRS validation models, household income was not included as a covariate due to high missingness within the variable. For each test, the environmental/lifestyle variables were modeled as follows: BMI, caffeine consumption, and PA were modeled as continuous variables, whereas T2D diagnosis, tobacco consumption, pesticide exposure, and head injury were modeled as binary variables. Models were of the form: PD status ~ age + sex + principal component 1 + principal component 2 + principal component 3 + principal component 4 + principal component 5 + environmental/lifestyle variable.

Next, we created 7 logistic regression models to test for the interaction between each of the 7 environmental variables and the effect of PRS. For each of the 7 variables, we modeled the effect of the lifestyle factor, the PRS (reflecting the OR associated with 1 SD increase in PRS), and the interaction between the factor and the PRS in the relevant age and sex-matched dataset. Each model followed the form: PD status ~ age + sex + principal component 1 + principal component 2 + principal component 3 + principal component 4 + principal component 5 + PRS + environmental/lifestyle variable + interaction between PRS and environmental/lifestyle variable. We then repeated these analyses, but substituted *GBA* carrier status for the PRS, to evaluate the interaction and effect of *GBA* carrier status. As we ran a total of 14 interaction tests, we used a Bonferroni adjusted p value of 3.6×10^{-3} (0.05/14 tests, 7 interaction models for PRS and 7 interaction models for *GBA*

TABLE 1. Characteristics of the Parkinson's Disease Cohort and Controls

Variable	Cases	N	Controls	N	*p*
Age, mean (SD), yr	73.1 (10.8)	18,819	73.0 (10.8)	545,751	1.9×10^{-1}
Female, n (%)	7,599 (40.2%)	18,819	303,979 (55.7%)	545,751	$<1 \times 10^{-300}$
PD duration, mean (SD)	6.8 (5.8)	18,819	N/A	N/A	N/A
PRS, mean (SD)	0.3 (1.0)	18,819	−0.01 (1.0)	545,751	$<1 \times 10^{-300}$
GBA carrier status, n (%)	1,438 (7.6%)	18,819	23,793 (4.5%)	545,751	1.3×10^{-80}
Physical activity, mean (SD)	3.1 (2.6)	14,695	3.1 (2.6)	602,495	2.6×10^{-1}
Q-norm body mass index, mean (SD), kg/m^2	−0.2 (1.1)	16,843	0.01 (1.0)	555,819	4.6×10^{-193}
Type 2 diabetes, n (%)	1,699 (11.5%)	16,425	64,090 (11.1%)	574,875	9.5×10^{-9}
Tobacco use, n (%)	6,547 (39.0%)	16,776	251,252 (46.8%)	536,832	1.9×10^{-106}
Pesticide exposure, n (%)	2,953 (42.1%)	7,022	88,186 (38.1%)	231,726	5.5×10^{-11}
Head injury, n (%)	3,137 (41.0%)	7,652	28,117 (33.4%)	84,172	2.4×10^{-22}
Caffeine consumption $\log_{10}(x + 75)$, mean (SD), mg	2.3 (0.3)	10,364	2.4 (0.4)	93,276	6.2×10^{-168}

Abbreviations: GBA = glucocerebrosidase; PD = Parkinson's disease; PRS = polygenic risk score.

carrier status) as the threshold for statistical significance for the interaction tests.

All statistical analyses were carried out using R v.4.1.2, and regressions were run using rlib23, a proprietary 23andMe software package.

Results

We constructed age- and sex-matched case–control datasets derived from the 23andMe research database (Methods; Table 1). The full dataset contained 18,819 PD cases (40.2% women) and 545,751 controls (55.7% women). At the time of data collection, PD cases and controls had an average age of 73.1 and 73.0 years, respectively ($SD_{PD} = 10.8$ years, $SD_{control} = 10.8$ years). The average years of PD duration for cases was 6.8 ($SD_{PD} = 5.8$ years). We tested 2 genetic factors: a PRS derived from the largest published PD GWAS (Nalls MA 2019) and *GBA* carrier status (PD = 7.6% carriers, control = 4.5% carriers). Table 2 shows the results of the 7 lifestyle traits and environmental exposures in the controls and PD cohort. Because data availability varied across self-reported variables, we constructed separate age and sex-matched datasets for each factor (median $N_{PD} = 14{,}692$, median $N_{control} = 520{,}056$). Age and sex distributions were similar across all datasets (PD mean age 71.5–72.9 years, PD percentage women 40.3–43.8%).

We observed negative associations between PD and caffeine intake (OR 0.43, 95% CI 0.41–0.46) tobacco use (OR 0.70, 95% CI 0.68–0.72), BMI (OR 0.79, 95% CI 0.78–0.80), and T2D (OR 0.86, 95% CI 0.82–0.91). We found positive associations between PD and pesticide exposure (OR 1.16, 95% CI 1.11–1.22), head injury (OR 1.31, 95% CI 1.25–1.38), PRS (OR per SD 1.41, 95% CI 1.39–1.43), and *GBA* carrier status (OR 1.73, 95% CI 1.63–1.83). There was no significant association between PD and physical activity (OR 0.99, 95% CI 0.99–1.00).

In the regression models with interaction terms, we found significant interactions between the PRS and T2D (OR 0.87, 95% CI 0.83–0.91, $p = 6.502 \times 10^{-8}$), BMI (OR 0.97, 95% CI 0.96–0.99, $p = 4.314 \times 10^{-4}$), PA (OR 1.01, 95% CI 1.01–1.02, $p = 8.745 \times 10^{-5}$), and tobacco use (OR 0.95, 95% CI 0.92–0.98, $p = 2.236 \times 10^{-3}$; Fig , Table 3). No significant interactions were observed in the *GBA* analysis.

To put the magnitude of these interactions in context, we calculated the crude prevalence of PD for each exposure in the top and bottom PRS quartiles. In the PRS-by-T2D interaction dataset, the crude prevalence of

TABLE 2. Regression Models Without an Interaction Effect

Variable	OR	95% CI	*p* value	N_{PD}	$N_{controls}$
Head injury	1.31	1.25–1.38	$*7.03 \times 10^{-28}$	7,652	84,172
Pesticide exposure	1.16	1.11–1.22	$*8.17 \times 10^{-10}$	7,022	231,726
Type 2 diabetes	0.86	0.82–0.91	$*1.97 \times 10^{-8}$	16,425	574,875
Q-norm BMI	0.79	0.78–0.80	$*2.48 \times 10^{-190}$	16,843	555,819
Tobacco use	0.70	0.68–0.72	$*2.42 \times 10^{-108}$	16,776	536,832
Caffeine consumption	0.43	0.41–0.46	$*1.80 \times 10^{-167}$	10,364	93,276
Physical activity	0.99	0.99–1.00	0.111	14,695	602,495

Note: Order of presentation: positive association, inverse association, null. *Significance was considered as $p < 0.05$.
Abbreviations: 95% CI = 95% confidence interval; BMI = body mass index; OR = odds ratio.

PD in the lowest PRS quartile was 7% higher in people with T2D (1.92%) compared with people without T2D (1.79%). However, in the highest PRS quartile, the prevalence of PD was 21% lower in people with T2D (3.37%) compared with people without T2D (4.17%).

In the PRS-by-BMI interaction dataset, the crude prevalence of PD in the lowest PRS quartile was 32% lower in people with BMI in the highest quartile (1.73%) compared with people with BMI in the lowest quartile (2.54%). In contrast, PD prevalence in the highest PRS quartile was 45% lower in people with BMI in the highest quartile (3.43%) compared with people with BMI in the lowest quartile (6.19%).

In the PRS-by-PA interaction dataset, the crude prevalence of PD in the lowest PRS quartile was 3% lower in people who were physically active >3.5 times per week (1.56%) compared with people who were physically active ≤1.5 times per week (1.60%). However, PD prevalence in the highest PRS quartile was 19% higher in people who were physically active >3.5 times per week (3.96%) compared with people who were physically active ≤1.5 times per week (3.32%).

In the PRS-by-smoking interaction dataset, the crude prevalence of PD in the lowest PRS quartile was 23% lower in ever smokers (1.72%) compared with never smokers (2.24%). In the highest PRS quartile, the prevalence of PD was 29% lower in ever smokers (3.76%) compared with never smokers (5.32%).

Discussion

In this large, cross-sectional case–control study, we observed evidence for statistical interactions between PRS and BMI, PRS and T2D, PRS and PA, and PRS and tobacco consumption on odds of PD.

There remains a large proportion of PD risk that continues to be unexplained by genetic variation, environmental exposures, lifestyle factors, or comorbidities. Exploring interactions between genetic and non-genetic factors may ultimately yield insights into disease risk through follow-up investigation. However, interpreting such interactions is far from straightforward, and so, rather than offering mechanistic insights, this study should be viewed as providing evidence for proof of principle.[25] We observed 4 significant statistical interactions in the present analysis. For 2 lifestyle variables, BMI and tobacco, the inverse associations between each variable and PD were attenuated in the presence of a higher genetic risk of PD, and magnified in the presence of a lower genetic risk.

For T2D, the nature of the interaction was more complicated, reversing directions of association at lower levels of PRS. The relationship between T2D and PD has been extensively studied, and there is increasing evidence for shared underlying pathways and mechanisms.[26] Most prospective cohort studies that ascertained T2D prior to PD diagnosis endorse a modest increase in the risk of PD associated with T2D.[27] However, cross-sectional and case–control studies often reveal inverse associations between T2D and PD, as seen in the present study. This raises the possibility of bias due to selective mortality, in study designs that are more susceptible to this type of bias.[27] Evidence for T2D conveying an increased risk of PD in high-quality, prospective studies is supported by Mendelian randomization.[27] We previously showed an interaction between T2D and PRS using data from UK Biobank.[28] In that prospective cohort study, the

interaction suggested that T2D was a more potent risk factor in those with a lower genetic liability toward PD. As such, in both studies, the magnitude of the association was greatest in those with lower genetic liability.

Drugs used to treat T2D are being widely repurposed and tested to see if they might modify the course of PD. Although the results presented here reflect PD risk rather than progression, they raise the possibility that

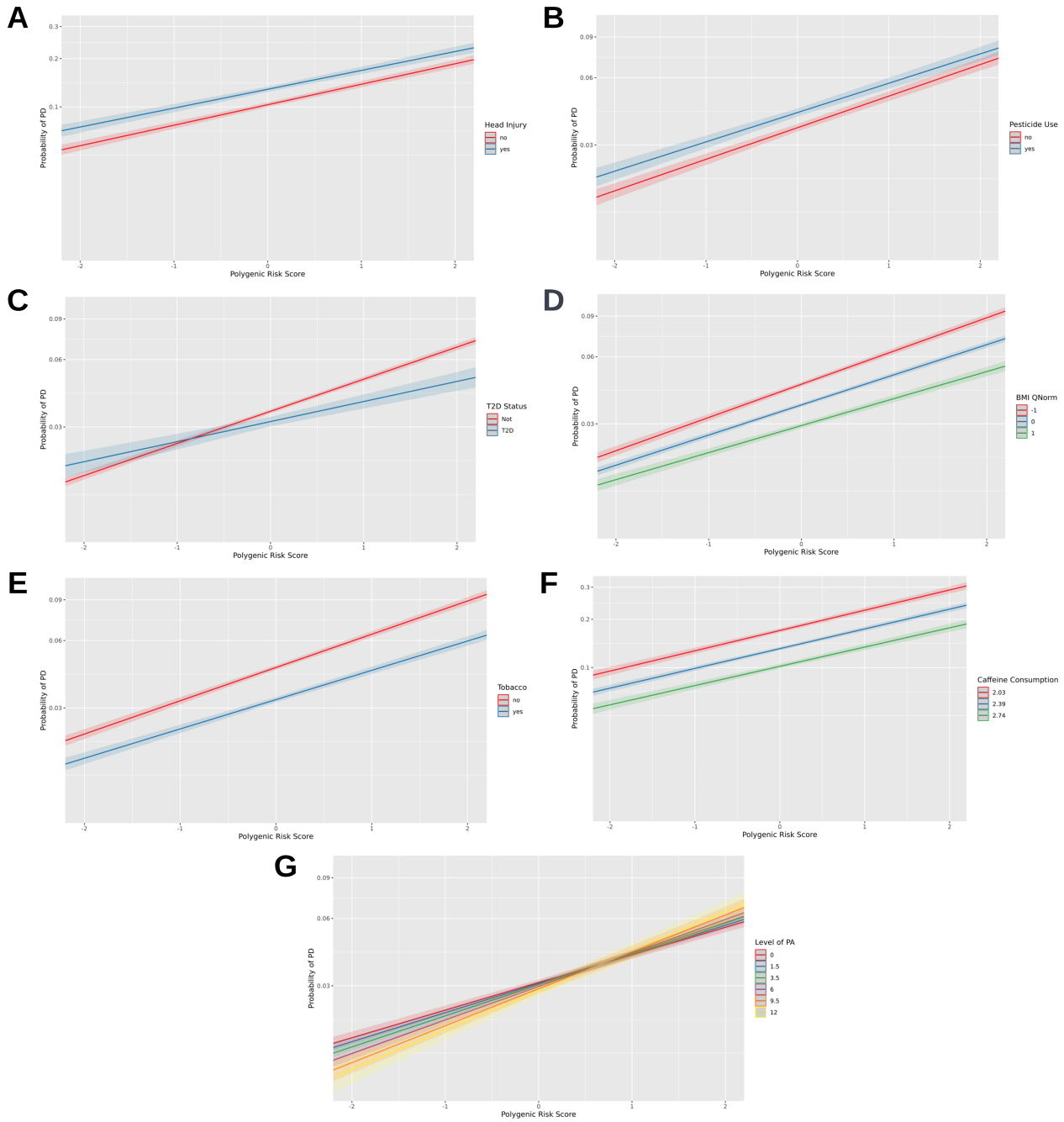

FIGURE: Marginal effects plots for the 7 polygenic risk score (PRS)-by-variable interaction models. Lines represent the fitted probability of Parkinson's disease (PD) for a given pair of PRS and variable values, shaded areas represent 95% confidence intervals. The *y*-axis shows the fitted probability of PD on a logit-transformed scale to preserve a linear relationship with the PRS and phenotypic factors. The *x*-axis window spans ±2 standard deviations of PRS, covering ~95% of individuals. The environment and lifestyle factor values plotted are as follows: (A) presence or absence of head injury; (B) presence or absence of pesticide exposure; (C) presence or absence of a type 2 diabetes (T2D) diagnosis; (D) quantile-normalized body mass index (BMI) at the sample mean (0), one standard deviation below the sample mean (−1), or one standard deviation above the sample mean (1); (E) presence or absence of tobacco use; (F) log10-transformed caffeine intake at the sample mean (2.39), one standard deviation below the sample mean (2.03), or one standard deviation above the sample mean (2.74); (G) number of 30-minute bouts of physical activity per week for all possible survey response options (0, 1.5, 3.5, 6, 9.5, or 12).

genetic stratification could be important when recruiting patients to such trials to identify subgroups that will have the best response.

There exists compelling observational evidence for an inverse association between PA and PD.[29–32] Again, a serious challenge is unpicking reverse causality from a causal relationship. Individuals who have undertaken regular PA appear to be at reduced risk of PD, but it is also probable that in the early stages of disease, PA reduces due to occult disease. We found no association between

TABLE 3. Regression Models With an Interaction Effect

Model		OR	95% CI	*p*	N
Head injury	PRS main effect	1.43	1.39–1.48	1.12×10^{-115}	91,140
	Variable main effect	1.33	1.26–1.40	5.49×10^{-28}	
	Interaction between PRS and variable	0.96	0.92–1.01	1.37×10^{-1}	
	Interaction between *GBA* and variable	1.08	0.89–1.30	4.22×10^{-1}	
Pesticide exposure	PRS main effect	1.42	1.37–1.46	1.36×10^{-108}	232,056
	Variable main effect	1.18	1.12–1.24	3.26×10^{-10}	
	Interaction between PRS and variable	0.96	0.92–1.01	1.13×10^{-1}	
	Interaction between *GBA* and variable	0.97	0.81–1.17	7.73×10^{-1}	
Type 2 diabetes	PRS main effect	1.42	1.40–1.44	$<1 \times 10^{-300}$	590,544
	Variable main effect	0.90	0.85–0.95	5.08×10^{-5}	
	Interaction between PRS and variable	0.87	0.83–0.91	$^{**}6.50 \times 10^{-8}$	
	Interaction between *GBA* and variable	0.81	0.65–0.99	4.83×10^{-2}	
Q-norm BMI	PRS main effect	1.39	1.37–1.41	$<1 \times 10^{-300}$	555,951
	Variable main effect	0.80	0.79–0.81	4.01×10^{-157}	
	Interaction between PRS and variable	0.97	0.96–0.99	$^{**}4.31 \times 10^{-4}$	
	Interaction between *GBA* and variable	0.95	0.89–1.00	6.15×10^{-2}	
Tobacco use	PRS main effect	1.44	1.41–1.47	6.28×10^{-288}	536,704
	Variable main effect	0.71	0.69–0.73	3.82×10^{-91}	
	Interaction between PRS and variable	0.95	0.92–0.98	$^{**}2.24 \times 10^{-3}$	
	Interaction between *GBA* and variable	0.82	0.72–0.93	1.46×10^{-3}	
Caffeine consumption	PRS main effect	1.50	1.31–1.73	8×10^{-9}	103,130
	Variable main effect	0.44	0.41–0.46	3.56×10^{-152}	
	Interaction between PRS and variable	0.97	0.91–1.03	2.85×10^{-1}	
	Interaction between *GBA* and variable	0.97	0.77–1.22	8.05×10^{-1}	
Physical activity	PRS main effect	1.34	1.30–1.37	4.62×10^{-111}	602,208
	Variable main effect	0.99	0.98–1.00	5.08×10^{-3}	
	Interaction between PRS and variable	1.01	1.01–1.02	$^{**}8.75 \times 10^{-5}$	
	Interaction between *GBA* and variable	1.02	1.00–1.05	5.69×10^{-2}	

Note: **Statistically significant interactions were considered as $p < 3.6 \times 10^{-3}$ (Bonferroni adjustment). Polygenic risk score reflects the OR associated with 1 SD increase in polygenic risk score.

Abbreviations: 95% CI = 95% confidence interval; OR = odds ratio; PRS = polygenic risk score; Q-norm BMI = quantile-normalized body mass index.

PA and PD in our regression model without an interaction effect. In our regression model with an interaction effect, however, a protective association was apparent with greater levels of PA in participants with lower genetic liability. Several PA-based intervention studies have already been conducted and have shown improvements in "off" state UPDRS scores, hinting at possible disease-modifying benefits.[33] Given the health benefits of PA, it is possible that survival bias contributed to the observed interaction between PA and PD genetic risk. For this to occur, mortality would have to be greater in people with (1) low levels of physical activity and high levels of PD genetic risk than in people with (2) the same low levels of physical activity, but low levels of PD genetic risk. Nonetheless, we cannot exclude the possibility of a detrimental effect of exercise beyond a certain age or stage of disease. For example, studies have shown that forced exercise following the acute phase of brain trauma can hinder synaptic transmission by promoting tissue sensitivity to stress responses.[34,35] PA-focused intervention studies are imminent for those at risk of future PD, and the current results suggest that genetic stratification could be important in their design and interpretation.

Observational studies examining the association between BMI and PD are complex to interpret given the dynamic nature of BMI during the course of PD. Generally, a reduction in BMI is seen following a diagnosis of PD and may precede the diagnosis.[36] This could be attributed to the disruption of normal homeostatic and hedonic mechanisms that result from neuroendocrine changes of disease pathogenesis.[37] Given the possibility of reverse causation, a meta-analysis of prospective cohort studies that measured BMI prior to PD diagnosis demonstrated no overall association.[38] Previous Mendelian randomization studies further examined the association of genetically estimated BMI and liability toward PD, again concluding an inverse association that was not thought to be explained by survival bias.[38,39]

The issue of survival bias is important when considering exposures that are associated with premature mortality and age-related outcomes, such as PD, including higher BMI and T2D or lower levels of physical activity. Several of the intriguing inverse associations that have been reported for PD might be driven, in part or entirely, by survivor bias or other types of bias. The cross-sectional nature of the present study prevents us from postulating a causal or clinically meaningful relationship between BMI and PD. Nonetheless, our study presents a novel finding, which is the possibility of this interaction to be modified according to genotype. Although individually such interactions may be of small magnitude, the significant shift and consequent risk, seen in population response, warrants further investigation.

An inverse association between smoking and PD risk has long been recognized. Epidemiological studies consistently highlight reduced odds of disease in individuals who smoke tobacco-containing cigarettes or are indirectly exposed to tobacco smoke.[11,40–43] In the present study, we not only observed the presence of an inverse association between tobacco consumption and odds of PD, but also a potential interaction between smoking and PRS, such that the protective association with smoking was greatest in those at higher genetic risk.

In this study, we also highlighted that pesticide exposure and a history of head injury are associated with a higher odds of PD, both of which have been well documented before.[10,44,45] Similarly, we replicated a well-known inverse association between caffeine and PD.[46–48] However, we did not find evidence of interaction with genotype for any of these associations.

Genetic stratification in future clinical trials seems inevitable, and will help underpin a general shift toward precision medicine.[49] Randomized controlled trials are an excellent way to examine causal relationships, but are not always ethical and/or practical. Based on the current results, existing and planned clinical trials focused on repurposed drugs for T2D and non-drug interventions, such as PA, may benefit from genetic stratification of participants to build on the initial observations we report here.[50,51]

A limitation of this study was that the definition of both PD cases and other related variables relied on self-reporting. This type of data collection can lead to bias and inaccuracy, but has been previously validated in terms of the variables reporting and the genetic data.[52,17] Furthermore, the dataset lacks temporal information regarding the 7 selected environment and lifestyle traits and PD, and is cross-sectional rather than longitudinal in nature. It should be noted that as data collection occurred on average 6.8 years after PD diagnosis, the associations observed could have been in part driven by survival bias and/or reverse causation. In addition, the 23andMe study population is not a random sample of the overall population, and results derived from this type of sampling may not be generalizable to individuals who are not well-represented. For example, our study only contained individuals of European descent. This means that the findings may not be generalizable and should be investigated in other ancestral groups.[53] Variability associated with ancestral diversity could account for differences in genetic risk factors, susceptibility to the environmental exposures, as well as interactions between the two.

The present study is the first to systematically examine interactions between selected environmental and lifestyle traits and common genetic variation related to PD. Use of 23andMe data meant that sufficiently large sample sizes could be used to investigate interaction, but the findings and implications should be followed by further research, including the re-analysis of data from previously completed T2D drug trials in PD and PA trials in PD, stratified by genotype.

Acknowledgements

We thank the research participants and employees of 23andMe for making this work possible. The following members of the 23andMe Research Team contributed to this study: Stella Aslibekyan, Adam Auton, Elizabeth Babalola, Robert K. Bell, Jessica Bielenberg, Jonathan Bowes, Katarzyna Bryc, Ninad S. Chaudhary, Daniella Coker, Sayantan Das, Emily DelloRusso, Sarah L. Elson, Nicholas Eriksson, Teresa Filshtein, Pierre Fontanillas, Will Freyman, Zach Fuller, Chris German, Julie M. Granka, Karl Heilbron, Alejandro Hernandez, Barry Hicks, David A. Hinds, Ethan M. Jewett, Yunxuan Jiang, Katelyn Kukar, Alan Kwong, Yanyu Liang, Keng-Han Lin, Bianca A. Llamas, Matthew H. McIntyre, Steven J. Micheletti, Meghan E. Moreno, Priyanka Nandakumar, Dominique T. Nguyen, Jared O'Connell, Aaron A. Petrakovitz, G. David Poznik, Alexandra Reynoso, Shubham Saini, Morgan Schumacher, Leah Selcer, Anjali J. Shastri, Janie F. Shelton, Jingchunzi Shi, Suyash Shringarpure, Qiaojuan Jane Su, Susana A. Tat, Vinh Tran, Joyce Y. Tung, Xin Wang, Wei Wang, Catherine H. Weldon, Peter Wilton, and Corinna D. Wong.

Author Contributions

A.J.N., B.M.J., K.H., and R.T. contributed to the conception and design of the study. A.R., J.S., K.H., L.K., and S.A. contributed to the acquisition and analysis of data. A.J.N., A.R., K.H., and R.T. contributed to drafting the text or preparing the figures.

Potential Conflicts of Interest

A.R., J.S., S.A., L.K., and K.H. are employed by and hold stock or stock options in 23andMe, Inc.

Data Availability

Consent for individual-level environmental and genetic data for 23andMe research participants has not been given for sharing.

References

1. Dorsey ER, Elbaz A, Nichols E, et al. Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol 2018;17:939–953.
2. Day JO, Mullin S. The genetics of Parkinson's disease and implications for clinical practice. Genes 2021;12:1006.
3. Fallin MD, Duggal P, Beaty TH. Genetic epidemiology and public health: the evolution from theory to technology. Am J Epidemiol 2016;183:387–393.
4. Nalls MA, Blauwendraat C, Vallerga CL, 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 2019;18:1091–1102.
5. Skrahina V, Gaber H, Vollstedt EJ, et al. The Rostock International Parkinson's disease (ROPAD) study: protocol and initial findings. Mov Disord 2021;36:1005–1010.
6. Gan-Or Z, Amshalom I, Kilarski LL, et al. Differential effects of severe vs mild GBA mutations on Parkinson disease. Neurology 2015;84:880–887.
7. Bandres-Ciga S, Diez-Fairen M, Kim JJ, Singleton AB. Genetics of Parkinson's disease: an introspection of its journey towards precision medicine. Neurobiol Dis 2020;137:104782.
8. Periñán MT, Brolin K, Bandres-Ciga S, et al. Effect modification between genes and environment and Parkinson's disease risk. Ann Neurol 2022;92:715–724.
9. Goldman SM, Tanner CM, Oakes D, et al. Head injury and Parkinson's disease risk in twins. Ann Neurol 2006;60:65–72.
10. Ascherio A, Chen H, Weisskopf MG, et al. Pesticide exposure and risk for Parkinson's disease. Ann Neurol 2006;60:197–203.
11. Noyce AJ, Bestwick JP, Silveira-Moriyama L, et al. Meta-analysis of early nonmotor features and risk factors for Parkinson disease. Ann Neurol 2012;72:893–901.
12. Jeong SM, Han K, Kim D, et al. Body mass index, diabetes, and the risk of Parkinson's disease. Mov Disord 2020;35:236–244.
13. Leehey M, Luo S, Sharma S, et al. Association of metabolic syndrome and change inUnified Parkinson's disease rating scale scores. Neurology 2017;89:1789–1794.
14. Lüth T, König IR, Grünewald A, et al. Age at onset of LRRK2 p. Gly2019Ser is related to environmental and lifestyle factors. Mov Disord 2020;35:1854–1858.
15. San Luciano M, Tanner CM, Meng C, et al. Nonsteroidal anti-inflammatory use and LRRK2 Parkinson's disease penetrance. Mov Disord 2020;35:1755–1764.
16. Jacobs BM, Belete D, Bestwick J, et al. Parkinson's disease determinants, prediction and gene–environment interactions in the UK biobank. J Neurol Neurosurg Psychiatry 2020;91:1046–1054.
17. Dorsey ER, Darwin KC, Mohammed S, et al. Virtual research visits and direct-to-consumer genetic testing in Parkinson's disease. Digit Health 2015:1–18.
18. Do CB, Tung JY, Dorfman E, et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet 2011;7:e1002141.
19. Heilbron K, Noyce AJ, Fontanillas P, et al. The Parkinson's phenome-traits associated with Parkinson's disease in a broadly phenotyped cohort. NPJ Parkinsons Dis 2019;5:4.
20. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084–1097.
21. Blauwendraat C, Reed X, Krohn L, et al. Genetic modifiers of risk and age at onset in GBA associated Parkinson's disease and Lewy body dementia. Brain 2020;143:234–248.

22. Alcalay RN, Levy OA, Waters CC, et al. Glucocerebrosidase activity in Parkinson's disease with and without GBA mutations. Brain 2015; 138:2648–2658.
23. Hartwig FP, Davies NM, Hemani G, Davey SG. Two-sample mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. Int J Epidemiol 2016; 45:1717–1726.
24. Durand EY, Do CB, Mountain JL, Michael MJ. Ancestry composition: a novel, efficient pipeline for ancestry deconvolution. bioRxiv 2014: 1–16. https://doi.org/10.1101/010512.
25. Clayton DG. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet 2009;5:e1000540.
26. Cheong JLY, de Pablo-Fernandez E, Foltynie T, Noyce AJ. The association between type 2 diabetes mellitus and Parkinson's disease. J Parkinsons Dis 2020;10:775–789.
27. Chohan H, Senkevich K, Patel RK, et al. Type 2 diabetes as a determinant of Parkinson's disease risk and progression. Mov Disord 2021;36:1420–1429.
28. Jacobs BM, Noyce A, Bestwick J, et al. Gene-environment interactions in multiple sclerosis: a UK biobank study. Neurol Neuroimmunol Neuroinflamm 2021;8:1–13.
29. Xu Q, Park Y, Huang X, et al. Physical activities and future risk of Parkinson disease. Neurology 2010;75:341–348.
30. Logroscino G, Sesso HD, Paffenbarger RS Jr, Lee I-M. Physical activity and risk of Parkinson's disease: a prospective cohort study. J Neurol Neurosurg Psychiatry 2006;77:1318–1322.
31. Sasco AJ, Paffenbarger RS Jr, Gendre I, Wing AL. The role of physical exercise in the occurrence of Parkinson's disease. Arch Neurol 1992;49:360–365.
32. Fang X, Han D, Cheng Q, et al. Association of Levels of physical activity with risk of Parkinson disease: a systematic review and meta-analysis. JAMA Netw Open 2018;1:e182421.
33. Van der Kolk NM, de Vries NM, Kessels RPC, et al. Effectiveness of home-based and remotely supervised aerobic exercise in Parkinson's disease: a double-blind, randomised controlled trial. Lancet Neurol 2019;18:998–1008.
34. Griesbach GS, Tio DL, Vincelli J, et al. Differential effects of voluntary and forced exercise on stress responses after traumatic brain injury. J Neurotrauma 2012;29:1426–1433.
35. Griesbach GS, Hovda DA, Tio DL, Taylor AN. Heightening of the stress response during the first weeks after a mild traumatic brain injury. Neuroscience 2011;178:147–158.
36. Simonet C, Bestwick J, Jitlal M, et al. Assessment of risk factors and early presentations of Parkinson disease in primary care in a diverse UK population. JAMA Neurol 2022;79:359–369.
37. De Pablo-Fernández E, Breen DP, Bouloux PM, et al. Neuroendocrine abnormalities in Parkinson's disease. J Neurol Neurosurg Psychiatry 2017;88:176–185.
38. Wang YL, Wang YT, Li JF, et al. Body mass index and risk of Parkinson's disease: a dose-response meta-analysis of prospective studies. PLoS One 2015;10:e0131778.
39. Heilbron K, Jensen MP, Bandres-Ciga S, et al. Unhealthy behaviours and risk of Parkinson's disease: a mendelian randomisation study. J Parkinsons Dis 2021;11:1981–1993.
40. Nielsen S, Gallagher LG, Lundin JI, et al. Environmental tobacco smoke and Parkinson's disease. Mov Disord 2012;27:293–297.
41. Sugita M, Izuno T, Tatemichi M, Otahara Y. March meta-analysis for epidemiologic studies on the relationship between smoking and Parkinson's disease. J Epidemiol 2001;11:87–94.
42. Ritz B, Ascherio A, Checkoway H, et al. Pooled analysis of tobacco use and risk of Parkinson disease. Arch Neurol 2007;64:990–997.
43. Thacker EL, O'Reilly EJ, Weisskopf MG, et al. Temporal relationship between cigarette smoking and risk of Parkinson disease. Neurology 2007;68:764–768.
44. Sherer TB, Richardson JR, Testa CM, et al. Mechanism of toxicity of pesticides acting at complex I: relevance to environmental etiologies of Parkinson's disease. J Neurochem 2007;100:1469–1479.
45. Jafari S, Etminan M, Aminzadeh F, Samii A. Head injury and risk of Parkinson disease: a systematic review and meta-analysis. Mov Disord 2013;28:1222–1229.
46. Bakshi R, Macklin EA, Hung AY, et al. Associations of lower caffeine intake and plasma urate levels with idiopathic Parkinson's disease in the harvard biomarkers study. J Parkinsons Dis 2020;10:505–510.
47. Saaksjarvi K, Knekt P, Rissanen H, et al. Prospective study of coffee consumption and risk of Parkinson's disease. Eur J Clin Nutr 2008; 62:908–915.
48. Luan Y, Ren X, Zheng W, et al. Chronic caffeine treatment protects against alpha-synucleinopathy by reestablishing autophagy activity in the mouse striatum. Front Neurosci 2018;12:301.
49. Leonard H, Blauwendraat C, Krohn L, et al. Genetic variability and potential effects on clinical trial outcomes: perspectives in Parkinson's disease. J Med Genet 2020;57:331–338.
50. Brauer R, Bhaskaran K, Chaturvedi N, et al. Glitazone treatment and incidence of Parkinson's disease among people with diabetes: a retrospective cohort study. PLoS Med 2015;12:e1001854.
51. Wahlqvist ML, Lee MS, Hsu CC, et al. Metformin-inclusive sulfonylurea therapy reduces the risk of Parkinson's disease occurring with type 2 diabetes in a Taiwanese population cohort. Parkinsonism Relat Disord 2012;18:753–758.
52. Munafò MR, Tilling K, Taylor AE, et al. Collider scope: when selection bias can substantially influence observed associations. Int J Epidemiol 2018;47:226–235.
53. Global Parkinson's Genetics Program. GP2: the global Parkinson's genetics program. Mov Disord 2021;36:842–851.