Machine Learning Modelling for Prospective Parkinson's Disease - the importance of inflammatory biomarkers and IGF1 - UKB study

Michael Allwright (Michael.allwright@sydney.edu.au)
The University of Sydney
Hamish Mundell
University of Sydney
Greg Sutherland
University of Sydney
Paul Austin
The University of Sydney
Boris Guennewig
The University of Sydney

Research Article

Keywords: PARKINSON'S, UK BIOBANK, IGF-1, MACHINE LEARNING, RISK SCORING, INFLAMMATION

Posted Date: March 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1410008/v1

Abstract

INTRODUCTION

We use the world-leading UK Biobank dataset (UKB) with over 500,000 participants and >10,000 variables across multiple data modalities to determine a ranking of candidate risk factors for Parkinson's Disease (PD) without *a priori* assumption.

METHODS

The Integrated Disease Explanation and Risk Scoring platform (IDEARS) applies machine learning algorithms to multi-modal personalised health data to determine individual disease risk and interpret the most important risk factors using mean SHAP score. IDEARS is applied to the UKB to determine the risk factors for PD.

RESULTS

IDEARS showed an improved discriminative performance (AUC=0.744) compared to a model using known risk factors (0.727). Age and gender had the highest mean SHAP scores. IGF-1, bilirubin, neutrophil/lymphocyte ratio (NLR, an inflammatory marker) and frailty factors were also ranked highly. A further investigation into IGF-1, bilirubin, AST:ALT and NLR showed elevated levels either in the period prior to diagnosis or at the point of diagnosis.

DISCUSSION

The IDEARS model outperformed an approach looking only at agreed risk factors despite making no *a priori* assumptions. Novel PD risk biomarkers including elevated IGF-1 and NLR are likely to play a role in disease mechanism. This panel of biomarkers may be used clinically to predict future PD risk, improve early diagnosis and to understand disease mechanism.

Introduction

Parkinson's Disease (PD) is the second most common neurodegenerative disease affecting over 6 million people worldwide and has seen a 3-fold increase in the last 30 years[1]. It is a movement disorder associated with a high level of disability for individual sufferers, and a great burden for care givers. To prospectively screen for PD and better elucidate the disease mechanism, there is a need to identify blood-borne biomarkers, as well as environmental and genetic factors that are associated with greater risk or are protective. This is critical because neurodegenerative processes in dopamine neurons of the midbrain start many years before PD diagnosis. Thus, there is a need for identification of future risk, enabling early interventions to be offered, which may take the form of beneficial lifestyle changes, or the development of novel neuroprotective agents.

Exposure to pesticides, consumption of dairy products, melanoma and traumatic brain injury are thought to increase the likelihood of PD diagnosis[2], while smoking, caffeine intake, high serum and urate concentrations, physical activity and the use of ibuprofen and other medications are considered protective[3,4]. There has also been some recent interest in the link between increased insulin like growth factor (IGF-1), bilirubin and inflammation, in the early phase of PD[5–8], whilst higher cholesterol levels are thought to cause a reduction in PD risk[9]. However, most studies which have examined PD risk to date consist of univariate hypothesis tests controlling for confounder variables of known risk factors. There have been few studies considering the associations from a wide range of variables together without *a priori* assumption. Employing machine learning to consider a full set of candidate risk factors together enables the significance of both established and novel risk factors to be evaluated side by side in an unbiased manner and, significantly, does not require any prior knowledge of PD risk factors.

The UK Biobank (UKB) is the largest deeply phenotyped epidemiological study in the world. A study has looked at the interaction between genetics and a set of established risk factors in predicting PD[10] and a separate study has confirmed the importance of anxiety, depression, family history of PD, excessive daytime sleepiness, pesticide exposure and being underweight, using logistic regression[11]. The association between lower lymphocyte count and PD has also been demonstrated in the UKB[12]. The methods applied in these studies were limited to logistic regression, Cox Proportional Hazards Survival Analysis and univariate approaches controlling for known risk factors. While these determine individual odds or hazard ratios associated with a small set of predictor variables, they neglect interaction and non-linear effects between variables, and do not have the capability to model more than a handful of variables. They therefore cannot take full advantage of the breadth of studies like the UKB. Cutting-edge machine learning techniques have already been applied to the UKB to determine risk factors for cardiovascular disease[13]. The ADNI dataset has been used with gradient boosting and SHAP to model APOE4[14], and neuroimaging data in those with mild cognitive impairment (MCI)[15].

IDEARS is an automated data processing, machine learning and visualisation platform that combines inpatient data, clinical assays and questionnaire data and applies feature engineering, classification models and feature importance methodologies to develop; an automated risk score and model performance metrics; the variables with the most significant associations with a disease using the novel kernel SHAP methodology to infer feature importance[15] and; visualization of disease risk profiles for specific variables. This study applies the IDEARS pipeline to determine individual PD risk for those aged 50–70 and validates this by comparing to a model derived from a set of well-established risk factors from previous studies. Through this process we present a novel ordered set of PD risk factors for consideration.

Materials And Methods

Ethics Approval

UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval. This approval means that researchers do not require separate ethical clearance and can operate under the RTB approval. All methods in this study were performed in accordance with the relevant guidelines and regulations of MREC.

Sample Selection

The UKB study recruited 502,253 subjects, aged 37–73 years in the United Kingdom between 2006–2010, performing a raft of clinical measurements and assays including clinical pathology screens, genotyping, neuroimaging and cognitive testing as well as medical information, health records and self-reported demographic and wellness data[16].

To establish our cohort, we considered participants who were aged 50–70 at baseline (the point at which they attended the assessment centre for their first set of tests). We excluded participants outside this age range due to the much lower risk of idiopathic PD prior to age 50 and the small numbers of participants aged over 70. We excluded those participants who were already diagnosed with PD at baseline or who developed PD within 2 years of baseline, to avoid skewing the predictions with the immediate pre-symptomatic phase of disease. We also excluded those who died within 10 years of baseline of something other than PD. This left a total of 384,591 participants in the cohort, 1,880 of whom developed PD within an average of 8.2 years from baseline.

IDEARS Platform

The IDEARS platform (Fig. 1) applies machine learning to health-related questionnaire data, longitudinal inpatient data (ICD10), blood assays, genetic and neuro-imaging data. It can be applied to any disease which can be categorized by a set of ICD10 codes and with a large enough number of cases (n > 1000). The full codebase can be accessed at https://github.com/binfnstats/ukb-IDEARS.

Data Integration Layer (Data Processing and Feature engineering)

Individuals who were diagnosed with either “Parkinson’s” or “Secondary Parkinsonism” between 2 and 10 years of their initial visit to the assessment centre were selected, according to the fields encoded by the UKB.

The most recent inpatient data (release date: September 2021) was used to identify the complete set of ICD10 codes corresponding to any condition for which the number of cases across our cohort exceeded 200 and for which a diagnosis was given prior to baseline. This resulted in 1,101 binary features corresponding to a participant having (1) or not having (0) a given disease at baseline. We derived a variable for the total number of conditions each participant had at baseline as well as the total number of conditions at within 20 illness groups defined through the ICD10 package in python.

Blood assay, clinical and self-reported questionnaire data were merged, and all variables which had greater than 80% non-missing observations were selected for the subsequent analysis. Feature engineering – which involved one-hot encoding and a conversion of variables to an ordinal score – was performed as part of the IDEARs pipeline.

A set of variables to represent risk factors with known associations to PD were developed – these included age, gender, neuroticism score, constipation, coffee intake, smoking status, exposure to pesticides, urban/rural living, depression, level of activity, a family history of PD, use at baseline of beta-blockers, ibuprofen and non-steroidal anti-inflammatories. Mean imputation of missing values was then performed on this processed dataset.

Risk Scoring and Model Explanation Layer

Two variable sets were used for modelling – V1 which included all variables extracted above (over 1500) and V2 which included only the consensus of risk factors derived from high-quality meta-analyses[4,3]. To avoid data leakage, data was first split into a training dataset, T1, and a holdout dataset, H1, with H1 containing a random sample of 20% of the full dataset, and 80% in T1.

Three classification model types were applied to T1 using a training dataset (70% random sample) and a test dataset (30% random sample). XGBoost (0.74) outperformed random forest (0.71) and Support Vector Machines (0.69) and was therefore selected. Due to XGBoost's inbuilt ability to handle unbalanced datasets, cases and controls were split in the ratio 1:20 and new controls were selected within each fold, thereby maximising the use of the control set. A ratio of 1:20 provides sufficient coverage of the full data but avoids the compute complexity of running on the entire dataset.

Hyper-parameter tuning was performed using the following hyperparameters: learning rate, minimum child weight, maximum depth, and positive weighting scale which determine the structure of the XGBoost algorithm. The set of hyperparameters which generated the highest AUC values for the model were selected to develop our model M1 which was then used in the subsequent analysis.

The full risk-scoring and model explanation layer was then resampled 50 times. This consisted of: 1) splitting data at random into a holdout dataset (H1–20% of records) and training dataset (T1–80% of records); 2) splitting the training data into a model data set (TT1, 70%) and TT2 (30%); 3) running model M1 on TT1 to select the top 25 features based on mean SHAP score calculated on TT2 for all cohorts (male, female and entire cohort); 4) incorporating these new candidate features alongside the known associations and applying M1 once more on these variable subsets, resampling 10 times on each cohort; and 5) evaluating the mean SHAP score of these models on the holdout dataset. Features which were selected in 1) more than 60% of the time and which had a mean SHAP score of 0.02 were then displayed by mean SHAP score, 6) applying model M1 to the same training dataset using only the V2 known association variables. The AUC metric was calculated on the holdout dataset H1 in each case to determine the discriminative performance of each model and to compare known associations (V2) model with the unbiased model (V1). Here we should include a sentence about why tt1 vs tt2 doesn't overfit.

Stratification by Gender and disease progression for key variables

The key variables output from the above analysis were manually grouped into ‘biometric’, ‘blood biomarkers’, ‘cardiovascular’, ‘demographic’, ‘frailty’ and ‘inflammation’. For the variables in these groups, we examined the data including those who already had been diagnosed with PD at baseline to determine how those variables changed in aggregate during PD progression as well as compared to those who would never develop PD (non-PD group). For both the male and female cohorts 2-sample t tests were performed to compare the means for each variable in the non-PD group with each disease stage, 5–10 years prior to disease diagnosis, 0–5 years prior to disease diagnosis and 0–5 years post disease diagnosis respectively.

Results

A cohort of 384,591 UKB participants, aged 50–70 at baseline met the inclusion criteria, of which 1,880 received a clinical diagnosis of Parkinson’s disease during the observation period (Table 1).

Table 1
Population Characteristics.

Variable	Cases	Controls	Total
n	1,880	382,711	384,591
Age at baseline (years)	63.6 +/- 4.5	60.1 +/- 5.4	60.1+/-5.4
Males	1,126 (0.64%)	174,722	175,898
Females	704 (0.34%)	207,989	208,693

The unbiased IDEARS model, applied to V1 (the full set of variables) had the best performance (AUC = 0.744), compared to V2, a comprehensive set of known risk factors identified from several high-quality meta-analyses (0.727). The AUROC graphs demonstrate superior performance of the IDEARS model in the total dataset, as well as when dividing the dataset by gender (Fig. 2). A statistically significant performance advantage of the unbiased IDEARS model is demonstrated based on a 100 resamples of the total dataset compared to known associations ($P > 0.0001$, Fig. 2B). The significant performance advantage is maintained with male and female datasets, with mean AUCs of 0.727 ($P > 0.0001$, Fig. 2D) and 0.719 ($P > 0.0001$, Fig. 2F), compared to 0.703 and 0.694 respectively.

Top features

The top features from the IDEARS model are shown in Fig. 3A. Variables from the meta-analyses which did not feature in our list of most important features, by average SHAP score were smoking status, traumatic brain injury and caffeine consumption. Higher age and male gender were the features with the

highest SHAP scores, in line with expectations given there known association with PD. IGF-1 (3rd), whether the participant was retired when they attended the assessment centre (4th), bilirubin levels (6th) and suffers from nerves (12th) were all associated with an increased risk of PD.

Features indicative of overall frailty were associated with greater risk of PD, with total ICD10 conditions at baseline (5th), number of treatments or medications (7th), hand grip strength (Left 10^{th,}) and usual walking pace (21st). A longer average duration to press a snap button was associated with increased PD risk (18th), suggesting cognitive decline could have some impact on PD. Features relating to inflammation were also important, with increased neutrophil/lymphocyte ratio (NLR) (13th) and neutrophil count (34th) being associated with PD, and increased C-reactive protein (17th) and lymphocyte count (11th) being protective. Cardiovascular and body fat variables appear to impact the risk of PD, with elevated total cholesterol being protective (14th) and larger waist circumference (9th) being associative.

The top features from the IDEARS model for in 1126 males and 704 females are shown in Figs. 3B&C, and the comparison of relative risk for each feature between males and females is shown in Fig. 3D. The removal of gender as a feature in the model leads to a relative increase in some features importance and the appearance of some additional features in the gender segregated lists. In males elevated IGF-1 (2nd) appeared to be a more important risk factor than in females (3rd), whereas for retired at baseline the opposite was true. Glycated haemoglobin (HbA1c), a marker of elevated blood sugar in the last 3 months, is 10th on the SHAP list for males and 12th for females. Alanine aminotransferase (ALT) was protective in both sexes (13th males, 27th females), and frailty-related features were mostly of equal importance in both sexes. AST:ALT ratio (4th), C-reactive protein (9th), neutrophil count (20th), having a parent with PD (22nd), and NLR (15th) were more associated with PD in males, whilst cholesterol (11th) and triglycerides (8th) were more protective of PD, suggesting inflammation, cardiovascular and genetic factors may be more important in males. In females, vitamin D (7th) and forced vital capacity (12th) were protective, whilst hip circumference (9th) and bilirubin (10th) were associated with increased PD risk.

IGF-1, bilirubin and AST:ALT ratio

Figure 4 shows IGF-1 and bilirubin levels, AST:ALT ratio and HbA1c in the 10 years preceding and 5 after a PD diagnosis in males and females compared to the non-PD group. IGF-1 was significantly elevated at both − 10 to -5 years (22.16 +/-5.74, P = 0.0002) and − 5 to 0 years before diagnosis (22.48+/-5.78, P = 0.0003) and 0–5 years after diagnosis in males (23.57+/-6.14, P > 0.0001) compared to non-PD (21.37+/-5.47, Fig. 4A). In females, IGF-1 was only significantly increased compared to non-PD (20.17+/-5.52) 0–5 years after diagnosis (21.63+/-6.20, P > 0.0001, Fig. 4B). Interestingly, when age normalising the IGF-1 levels in females between non-PD and PD, there is an emergence of a significant increase in IGF-levels − 10 to -5 years before diagnosis (P = 0.000362, data not shown). This may suggest that IGF-1 levels are elevated in females who later develop PD, but maybe obscured by other age-related effects of IGF-1, however clearly the positive association is weaker than in males.

Bilirubin was significantly elevated at -5 to 0 years before diagnosis (11.23+/-5.43, P = 0.0003) and 0–5 years after diagnosis (11.14+/-5.87, P = 0.0014) in males compared to non-PD (10.25+/-4.78, Fig. 4C). Despite having a greater SHAP score in females, bilirubin was not significantly elevated compared to non-PD (8.05+/-3.53) following diagnosis (8.53+/-4.51, P = 0.0885, Fig. 4D). AST:ALT ratio was significantly elevated at both − 10 to -5 years (1.23 +/-0.40, P = 0.0003) and − 5 to 0 years before diagnosis (1.25 +/-0.37, P > 0.0001) and 0–5 years after diagnosis in males (1.27 +/-0.45, P > 0.0001) compared to non-PD (1.17 +/-0.40, Fig. 4E). In females, the AST:ALT ratio was only significant increased compared to non-PD (1.34 +/-0.43) 0–5 years after diagnosis (1.45 +/-0.63, P = 0.0008, Fig. 4F). Therefore, IGF-1 and bilirubin levels, and the AST:ALT ratio are all elevated in males prior to diagnosis, whereas in females they increase around the time of diagnosis. According to the SHAP chart HbA1c showed a protective association with the risk of PD. However, looking at the values over time, HbA1c was significantly elevated in males 10 to 5 years before diagnosis (37.88 +/-7.76, P < 0.001) compared to non-PD (37.04 +/-8.65), and 5 to 0 years before diagnosis (37.70 +/-7.83, P < 0.001) in females compared to non-PD (36.67 +/-6.03). At the time of diagnosis there was no significant difference in HbA1c in either sex. Interestingly, the threshold for diagnosis of type 2 diabetes is HbA1c > 48 mmol/mol [18], and this equates to 5.7% of the PD group and 3.5% of the non-PD group, suggesting that although HbA1c is not significantly elevated at diagnosis there may be a greater prevalence of type 2 diabetes.

Figure 5 shows inflammatory variables in the 10 years preceding and 5 after a PD diagnosis in males and females compared to the non-PD group. Neutrophil count was significantly elevated − 10 to -5 years (4.46 +/-1.42, P = 0.0152) and 0–5 years after diagnosis in males (4.64 +/-1.42, P = 0.0002) compared to non-PD (4.33 +/-1.44, Fig. 5A). In females, neutrophil count was not significantly increased compared to non-PD at any timepoint. Lymphocyte count was significantly reduced at 10 − 5 years before diagnosis (1.76 +/-0.58, P = 0.0087) and 0–5 years after diagnosis (1.71 +/-0.68, P = 0.0152) in males compared to non-PD (1.91 +/-1.46, Fig. 5C). In females, lymphocyte count was significantly decreased compared to non-PD (2.04 +/-1.12) only after diagnosis (1.80 +/-0.51, P = 0.0885, Fig. 5D).

NLR, a common marker of stress and inflammation, was significantly elevated − 10 to -5 years (2.77 +/-1.32, P < 0.0001) and 0–5 years after diagnosis in males (3.03 +/-1.39, P > 0.0001) compared to non-PD (2.53 +/-1.47, Fig. 5E), with a similar pattern in females; -10 to -5 years (2.33 +/-0.92, P = 0.0491), 0–5 years after diagnosis (2.55 +/-1.05, P = 0.0002 and non-PD (2.21 +/-1.13, Fig. 5F). The relative risk of PD increases significantly with an elevated NLR, but that risk is reduced when you compare those that were taking ibuprofen at baseline to those that weren't (Fig. 6). Whilst ibuprofen shows a strong protective effect at all NLRs, it is most apparent in those with the highest ratio, suggesting inflammation plays an important role in the disease mechanism.

According to the IDEARS model C-reactive protein, demonstrated a protective relationship with PD with a relatively high SHAP score (8th in males, 16th in females), however there were no timepoints that were significantly decreased compared to non-PD in either sex (Fig. 5G&H).

In summary, inflammatory features in males were most consistently different to non-PD at -10 to -5 years diagnosis and after diagnosis, whereas the only predictive inflammatory biomarker in females was the NLR which was increased at -10 to -5 years before, and 0 to 5 years after diagnosis. An interesting observation in the inflammatory features is that there were no significant differences at -5 to 0 years before diagnosis in either sex, which could suggest a return to towards baseline at this pre-symptomatic period, and a second increase after diagnosis. Therefore, the NLR appears to be the most consistent inflammatory biomarker associated with PD, and ibuprofen consumption appears to mitigate the negative effects of an elevated an NLR ratio.

Frailty

The IDEARS model revealed several frailty-related variables as being associated with the development of PD and these are shown in Fig. 7 in the years preceding and following diagnosis. Total ICD10 diagnoses were significantly increased at -10 to -5 years (3.88 +/-5.18, male, p = 0.046; 4.07 +/-5.42, female, p = 0.0074), -5 to 0 years (4.50 +/-5.68, male, P > 0.0001; 4.50 +/-6.338, female, p = 0.0074) and 0–5 years after diagnosis (4.495 +/-6.338, male, p = 0.046; 4.52 +/-5.66, female, p = 0.0074) compared to non-PD (2.78 +/-4.34, male; 2.75 +/-4.03, female, all comparisons P > 0.0001, Fig. 7A&B).

Total treatment/medications was similarly correlated with a greater risk of PD, being significantly increased at -10 to -5 years (3.46 +/-3.19, male; 3.77 +/-3.24, female), -5 to 0 years (3.77 +/-3.24, male; 4.28 +/-3.53, female) and 0–5 years after diagnosis (4.73 +/-3.64, male; 4.73 +/-3.64, female) compared to non-PD (2.63 +/-2.79, male; 2.79 +/-2.78, female, all comparisons P > 0.0001, Fig. 7C&D).

Considering that PD is a movement disorder, it was unsurprising that grip strength in both hands (Fig. 8E-H), and usual walking speed (see supplementary file) were significantly reduced for both sexes at all timepoints (P > 0.0001 for all comparisons). Perhaps the most interesting observation is that these decreases are apparent at up to 10 years before diagnosis occurs.

Greater forced vital capacity was demonstrated be protective particularly in females, however a significant reduction was seen in both sexes. Forced vital capacity was significantly reduced at -10 to -5 years (4.15 +/-0.93, male, p = 0.0025; 2.86 +/-0.63, female, p < 0.0001), -5 to 0 years (4.038 +/-0.807, male, p < 0.0001; 2.765 +/-0.65, female, p < 0.0001) and 0–5 years after diagnosis (2.77 +/-0.65, male, p = 0.0139; 2.82 +/-0.57, female, p < 0.0001) compared to non-PD (4.26 +/-0.94, male; 3.03 +/-0.70, female, Fig. 7I&J). Overall, the frailty-related features show very strong associations with PD risk in both sexes even from 10 years prior to diagnosis.

Cardiovascular features and body adiposity

Figure 8 shows a range of cardiovascular features at timepoints before and after PD diagnosis. Decreased total cholesterol correlated with a lower risk of PD in both sexes, cholesterol was significantly reduced at both − 10 to -5 years (5.17 +/-1.11, P < 0.0001), -5 to 0 years (5.15 +/-1.1, P < 0.0001) and 0–5 years after diagnosis in males (5.01 +/-1.04, P < 0.0001) compared to non-PD (5.44 +/-1.14, Fig. 8A). In

females, cholesterol was significantly decreased compared to non-PD (6.022 +/-1.133) only at -5 to 0 years (5.779 +/-1.129, P = 0.0012). Given the protective effect of total cholesterol we investigated HDL and LDL cholesterol. For HDL there were no timepoints that were significantly different compared to non-PD, and levels were consistent across all groups (Fig. 8C&D). However, a reduction in LDL was apparent a multiple timepoints for both sexes. LDL was significantly reduced at -10 to 5 years (3.23 +/-0.84, P < 0.0001) and − 5 to 0 years before diagnosis (3.23 +/-0.84, P < 0.0001) and 0–5 years after diagnosis (3.12 +/-0.81, P < 0.0001) in males compared to non-PD (3.44 +/-0.87, Fig. 8E). In females, LDL was significantly decreased compared to non-PD (3.73 +/-0.88) only at -5 to 0 years (3.54 +/-0.85 P = 0.0015, Fig. 8F). Interestingly, waist circumference was significantly increased in both sexes compared to non-PD (97.61 +/-11.32, male; 85.42 +/-12.45, female) but only at -10 to -5 years before diagnosis (98.44 +/-10.93, male, p = 0.046; 87.053 +/-12.137, female, p = 0.0074, Fig. 8G&H). In summary, total cholesterol and LDL appear to be protective, this indicates people with a reduced risk of heart disease a more likely to developed PD, however increased central adiposity, indicated by increased waist circumference (and hip circumference in females) is associated with increased PD risk in the next 10 years.

Discussion

We presented the IDEARS platform, which uses state-of-the-art machine learning algorithms XGBoost and SHAP to provide a ranking of risk factors for PD using the world's largest and most comprehensive prospective community study, the UK Biobank. As expected, ageing was by far the most significant factor in predicting PD followed by gender, PD is far more prevalent in males, a fact indicated by the larger number of males with PD in the UKB. Our unbiased machine learning approach uncovered a novel set of features most associated with PD. Interestingly, several well-established risk factors thought to have a high level of association with PD were not identified in the most important features in our model (e.g., pesticide exposure, smoking status, traumatic brain injury and caffeine consumption).

Of note is the importance of insulin-like growth factor 1 (IGF-1), which presented in the top 4 most important features, based on mean SHAP score in the combined dataset, and male and female lists. On deeper inspection of the data, it was clear that IGF-1 levels were elevated in males up to 10 years before disease onset, and when age normalising the data this patten could also be unearthed in females. IGF-1 is an endocrine, paracrine and autocrine hormone that is a primary mediator of the effects of growth hormone. Major functions of IGF- include insulinlike activity, cell proliferation and survival, antioxidant effects and neuroprotection. *In vivo* studies have demonstrated IGF-1 deficiency results in increased oxidative stress, inflammation, neuronal cell death and cognitive deficits that can be improved by exogenous IGF-1[19,20]. It is well documented that IGF-1 is elevated in serum at diagnosis in PD patients, and levels at this time correlate with disease severity[5,8]. To account for the discrepancy in the beneficial effects of IGF-1 and the fact it is increased in PD, it has been hypothesised that IGF-1 signalling is defective in PD, resulting in a decrease in the neuroprotective effects and reduction in the brains ability to buffer oxidative damage. Moreover, IGF-1 signalling is known to be dysregulated by both toxin-induced inflammation and central obesity[5, 21, 22], which is consistent with our model identifying prospective

biomarkers predictive of greater PD risk in these categories. Therefore higher-than-average IGF-1 levels, especially in men, years before diagnosis may be indicative of a compensatory mechanism in response to dysregulated IGF-1 signalling. Our findings suggest that IGF-1 should be further considered as a prognostic biomarker for PD risk.

The IDEARS model identified bilirubin levels as being elevated in the 5 years before diagnosis but only in males, although the was trend for an increase in females. There is evidence that bilirubin levels are elevated in the early years post PD diagnosis, but to our knowledge no one has investigated the higher levels pre-PD diagnosis. A recent meta-analysis concluded that there was an increase in total bilirubin serum levels in PD patients, however it was more robust in the Caucasian population[6], consistent with the UKB cohort. Furthermore, bilirubin levels negatively correlated with disease severity[23] and dopamine replacement therapy elevated bilirubin levels compared to drug naïve PD patients[24]. Bilirubin is part of the heme oxygenase antioxidant pathway, therefore elevated levels in PD are likely a compensatory mechanism to increased oxidate stress in the parkinsonian brain. Thus, like IGF-1, bilirubin should be considered as a prognostic biomarker for PD risk, however it may not be a strong marker in females and some ethnicities.

AST:ALT was elevated in males up to 10 years before PD diagnosis but only increases in females after diagnosis, this is consistent with elevated ALT being protective, and being higher in the male SHAP list. Elevated AST:ALT ratios between 1–2 are indicative of non-alcoholic fatty liver disease (NAFLD) or non-alcoholic steatohepatitis (NASH), whilst levels < 2 are indicative of alcoholic liver disease[25,26], therefore the moderate increases in the UKB PD cohort may be indicative of NAFLD/NASH, although some individuals in the PD group have levels above 2. A recent study of NAFLD and PD found that there was greater risk of PD in females with NAFLD[27], and an earlier study found that NASH in males and females with hepatitis B and C infection led to a greater PD risk[28]. With that said, NAFLD is associated with cardiovascular disease and metabolic disorders which does not fully align with our other findings (see below)[29]. Whilst more research on NAFLD and PD is required, our findings indicate elevated AST:ALT may be a useful prospective biomarker of PD, especially in males.

The IDEARS model identified several features associated with cardiovascular health and body adiposity. Total and LDL cholesterol levels were reduced in PD in males 10 years before diagnosis but only 5 years in females. This observation is in keeping with a large population-based study of 261,638 statin-free individuals, which identified that males who had lower levels of total and LDL cholesterol were at a greater risk of developing PD, however there was no significant differences in females[9]. Given lower LDL levels, PD patients have shown a reduced risk of myocardial infarction and stroke[30,31], and it has been hypothesised that the reduced cholesterol levels may be due to nonmotor peripheral symptoms, such as constipation, that can manifest before motor symptoms appear[9].

Cardiovascular health is also strongly linked to metabolic regulation, and there are mixed findings on the co-morbidity of type 2 diabetes and PD, with some studies showing an increase [32], and others showing a

reduced prevalence[31,33]. As mentioned above HbA1c is higher several years before PD onset, but not elevated at the time of diagnosis, although the proportion of the PD group with HbA1c in the diabetic range is slightly higher than the non-PD group. Therefore, further research is needed to investigate the possible associations of diabetes and PD.

Several epidemiological studies have linked central adiposity to PD[34,35], which is consistent with output from the IDEARS model with increased waist circumference 10 years before diagnosis a risk factor in both sexes (and hip circumference in females). Although this observation may be at odds with better cardiovascular and metabolic health in general, body fat distribution is likely key factor, and increased adiposity has also been hypothesised to modulate IGF-1 signalling[5,21]. Clearly, more research is required to better understand the complex interactions of body adiposity and the risk of PD.

Several features relating to the immune system were identified by the IDEARS model, specifically an increase in neutrophil count, a decrease in lymphocyte count and in NLR, were all identified to be altered both 10 years before and at diagnosis in males, whilst only NLR followed the same pattern in females. An elevated neutrophil count is associated with the occurrence, progression and severity of inflammation or infection, whereas a decreased lymphocyte count, as part of the adaptive immune response, is heavily depressed by stress. Thus, NLR is considered a compound biomarker of inflammation and stress, and therefore it is perhaps not surprisingly that NLR is the most robust and consistent example of a prospective biomarker of PD risk from the IDEARS model. A recent study demonstrated similar findings with increased NLR in 100 PD patients, but no change in Alzheimer's disease [7]. Increased neutrophil count and NLR are in keeping with the literature that inflammation and infection are risk factors for PD. NLR may therefore be considered a useful prospective biomarker for the risk of PD, however as it is associated with many other chronic diseases, it should be used in combination with other biomarkers identified by our model.

Epidemiological studies have revealed viral (e.g. *influenza*, *HSV, hepatitis*) and bacterial (e.g. *C. pneumonia and H. pylori*) infections are associated with an increased risk of developing PD [28,36–39]. Inflammatory conditions, such as head trauma, allergic rhinitis and exaggerated allergic reactions following insect stings, have been linked to an increased risk of developing PD[40–43]. Neuroinflammation is also a common pathological hallmark seen in the PD brain[44–47]. Conversely, long-term use of non-steroidal anti-inflammatory drugs (NSAIDs) reduce the risk of developing PD [48–51]. Our analysis clearly demonstrates a protective effect of Ibuprofen use in the UKB participants, which was more pronounced in at higher NLR.

The reduction in lymphocyte count well before PD in our study is consistent with another recent analysis using the UKB dataset (thus validating our approach)[12], as well as a meta-analysis that showed decreased numbers of $CD3^+$ and $CD4^+$ lymphocyte subsets in intermediate and late-stage PD, whilst a decrease in $CD8^+$ T lymphocytes was also observed[52]. It is interesting to observe that this reduction in lymphocyte count occurs up to 10 years prior to diagnosis in males, and therefore maybe a better

prospective marker in men. It is noteworthy that 'suffers from nerve' was also a highly ranked risk factor in the IDEARS model (8th overall), and therefore the PD group may have higher-than-average stress levels, which could depress lymphocyte counts. More detailed analyses of $CD4^+$ T lymphocyte subsets suggests that they are skewed towards proinflammatory phenotypes (i.e., increased Th1, Th17, and reduced Th2 and Tregs) in PD patients[53–55]. The inflammatory milieu in PD is a likely contributor to decreased IGF-1 signalling mentioned previously[5,19,20]. Overall, these findings imply a predisposition to PD may be established by conditions that induce peripheral inflammation (injury/infection) and stress, or in individuals with an immune system skewed towards inflammation.

Given that PD is an age-related motor disease it was unsurprising that the IDEARS model identified several features associated with frailty and cognitive function. Reduced hand grip strength and decreased walking pace can be considered early markers of motor dysfunction and given they are significantly reduced in both sexes 10 years before diagnosis they should be considered as useful clinical measures to predict the risk of PD onset. Existing literature has identified the importance of these factors. Hand grip strength and reduced dexterity have been reported as a predictors of motor symptom severity in PD[56]. Slow walking speed has been correlated with both advanced aged and PD severity[57], and it is also one of the first complaints in the early stages of the disease[58]. Increased number of ICD conditions at baseline, increased number of medications/treatments taken and reduced forced vital capacity were also apparent 10 years before diagnosis in both sexes, they are likely indictors of general ill health and multiple co-morbidities in PD patients. Arthritis, hypertension, atrial fibrillation, depression, back problems, and cataracts are commonly reported co-morbidities of PD [32,59], and require a wide range of treatments.

Other significant gender differences were observed with parental PD being more important for men than women, which may suggest that since idiopathic PD has a phenotype that strongly overlaps with monogenic forms of the disease[60], there may be a greater genetic component in idiopathic PD in males. While platelet count, vitamin D and testosterone (normalised by gender) more important for women. Vitamin deficiency has been linked to neurodegenerative diseases, and a deficiency in vitamin D in particular has been linked to reduced dopamine levels and alpha-synuclein accumulating, which are pathological hallmarks of PD [61]. Vitamin D has been shown to have neuroprotective, anti-inflammatory and antioxidant effects *in vitro*[62], however a recent metanalysis could not conclude clear benefits of vitamin D supplementation in reducing PD risk[63].

The application of a novel methodology in the IDEARs pipeline has enabled us to examine a much larger range of variables without *a priori* assumption. The advantages of using XGBoost and SHAP in this context is in the ability to consider a large number of variables and accurately determine their importance in the model while implicitly modelling interactions between variables, resulting in a demonstratively higher AUC. The disadvantage is the black box nature of this approach. We have sought to mitigate this by providing a separate univariate analysis of individual variables. In addition to the power of determining the most significant risk factors in driving PD, this approach could be used separately to provide a risk score which would be more accurate than existing methods.

Conclusion

In summary our novel unbiased IDEARS model identified a novel set of risk factors for PD that diverge considerably from the most well-established risk factors thought to have a high level of association with PD. The most promising biomarkers for PD risk are elevated IGF-1, bilirubin, AST:ALT, NLR and reduced total and LDL cholesterol. Nearly all these biomarkers demonstrated a consistent change before PD onset in males, however this pattern was not always as robust in females (although from a smaller sample size). Given the non-specific nature of some of these biomarkers (e.g. AST:ALT, NLR), we suggest that they would be best used in combination to predict PD risk. If the hoped-for development of neuroprotective treatments for PD is fruitful, our biomarker panel may help identify those at heightened risk who may benefit most from prophylactic treatment. Features indicative of frailty, particularly those that relate to motor dysfunction, such as walking speed and hand grip strength, as well as a high number of co-morbidities, were strongly associated with increased PD risk in both sexes, and these signs could serve as useful clinical indications leading to earlier diagnosis.

Declarations

Data Availability Statement

The datasets generated in the current study are included in this manuscript and its supplementary files. The raw data used during the current study are not publicly available due to this being private data from the UK Biobank but are available from the UK Biobank directly https://www.ukbiobank.ac.uk/ upon application. Further data breakdowns can be provided by the corresponding author on reasonable request.

References

1. Dorsey, E. R.*et al.* Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology* **17**, 939-953, doi:10.1016/s1474-4422(18)30295-3 (2018).
2. Noyce, A. J.*et al.* Meta-analysis of early nonmotor features and risk factors for Parkinson disease. *Annals of Neurology* **72**, 893-901, doi:10.1002/ana.23687 (2012).
3. Ascherio, A. & Schwarzschild, M. A. The epidemiology of Parkinson's disease: risk factors and prevention. *Lancet Neurol* **15**, 1257-1272, doi:10.1016/S1474-4422(16)30230-7 (2016).
4. Bellou, V., Belbasis, L., Tzoulaki, I., Evangelou, E. & Ioannidis, J. P. A. Environmental risk factors and Parkinson's disease: An umbrella review of meta-analyses. *Parkinsonism & Related Disorders* **23**, 1-9, doi:10.1016/j.parkreldis.2015.12.008 (2016).
5. Castilla-Cortázar, I., Aguirre, G. A., Femat-Roldán, G., Martín-Estal, I. & Espinosa, L. Is insulin-like growth factor-1 involved in Parkinson’s disease development? *Journal of Translational Medicine* **18**, doi:10.1186/s12967-020-02223-0 (2020).

6. Jin, J.-N., Liu, X., Li, M.-J., Bai, X.-L. & Xie, A.-M. Association between serum bilirubin concentration and Parkinson's disease: a meta-analysis. *Chinese Medical Journal* **134**, 655-661, doi:10.1097/cm9.0000000000001300 (2021).
7. Kara, S. P., Altunan, B. & Unal, A. Investigation of the peripheral inflammation (neutrophil-lymphocyte ratio) in two neurodegenerative diseases of the central nervous system. *Neurol Sci*, 1-9, doi:10.1007/s10072-021-05507-5 (2021).
8. Li, D.-H., He, Y.-C., Quinn, T. J. & Liu, J. Serum Insulin-Like Growth Factor-1 in Patients with De Novo, Drug Naïve Parkinson's Disease: A Meta-Analysis. *PLOS ONE* **10**, e0144755, doi:10.1371/journal.pone.0144755 (2015).
9. Rozani, V.*et al.* Higher serum cholesterol and decreased Parkinson's disease risk: A statin-free cohort study. *Movement Disorders* **33**, 1298-1305, doi:10.1002/mds.27413 (2018).
10. 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, doi:10.1136/jnnp-2020-323646 (2020).
11. Belete, D., Jacobs, B., Schrag, A. & Noyce, A. Exploring the Parkinson's disease phenome in the UK Biobank population (4040). *Neurology* **94**, 4040 (2020).
12. Jensen, M. P.*et al.* Lower Lymphocyte Count is Associated With Increased Risk of Parkinson's Disease. *Annals of Neurology* **89**, 803-812, doi:10.1002/ana.26034 (2021).
13. Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F. & van der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS One* **14**, e0213653, doi:10.1371/journal.pone.0213653 (2019).
14. Petersen, R. C.*et al.* Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology* **74**, 201-209, doi:10.1212/wnl.0b013e3181cb3e25 (2010).
15. <Bloch_SHApley_ADNI_2021.pdf>. doi:10.1186/s13195-021-00879-4.
16. 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, doi:10.1371/journal.pmed.1001779 (2015).
17. Livingston, G.*et al.* Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet* **396**, 413-446, doi:10.1016/s0140-6736(20)30367-6 (2020).
18. d'Emden, M. Glycated haemoglobin for the diagnosis of diabetes. *Australian Prescriber*, 98-100, doi:-10.18773/austprescr.2014.037 (2014).
19. Castilla-Cortazar, I.*et al.* An experimental model of partial insulin-like growth factor-1 deficiency in mice. *J Physiol Biochem* **70**, 129-139, doi:10.1007/s13105-013-0287-y (2014).
20. Puche, J. E., Muñoz, Ú., García-Magariño, M., Sádaba, M. C. & Castilla-Cortázar, I. Partial IGF-1 deficiency induces brain oxidative damage and edema, which are ameliorated by replacement therapy. *Biofactors* **42**, 60-79, doi:10.1002/biof.1255 (2016).

21. Aguirre, G. A., De Ita, J. R., de la Garza, R. G. & Castilla-Cortazar, I. Insulin-like growth factor-1 deficiency and metabolic syndrome. *Journal of Translational Medicine* **14**, 3, doi:10.1186/s12967-015-0762-z (2016).
22. Aguirre, G. A., González-Guerra, J. L., Espinosa, L. & Castilla-Cortazar, I. Insulin-Like Growth Factor 1 in the Cardiovascular System. *Rev Physiol Biochem Pharmacol* **175**, 1-45, doi:10.1007/112_2017_8 (2018).
23. Moccia, M.*et al.* Increased bilirubin levels in de novo Parkinson's disease. *Eur J Neurol* **22**, 954-959, doi:10.1111/ene.12688 (2015).
24. Scigliano, G.*et al.* Increased plasma bilirubin in Parkinson patients on L-dopa: evidence against the free radical hypothesis? *Ital J Neurol Sci* **18**, 69-72, doi:10.1007/bf01999565 (1997).
25. Hall, P. & Cash, J. What is the real function of the liver 'function' tests? *Ulster Med J* **81**, 30-36 (2012).
26. Sorbi, D., Boynton, J. & Lindor, K. D. The ratio of aspartate aminotransferase to alanine aminotransferase: potential value in differentiating nonalcoholic steatohepatitis from alcoholic liver disease. *Am J Gastroenterol* **94**, 1018-1022, doi:10.1111/j.1572-0241.1999.01006.x (1999).
27. Jeong, S. M.*et al.* Sex differences in the association between nonalcoholic fatty liver disease and Parkinson's disease. *Parkinsonism Relat Disord* **93**, 19-26, doi:10.1016/j.parkreldis.2021.10.030 (2021).
28. Goldstein, L., Fogel-Grinvald, H. & Steiner, I. Hepatitis B and C virus infection as a risk factor for Parkinson's disease in Israel-A nationwide cohort study. *J Neurol Sci* **398**, 138-141, doi:10.1016/j.jns.2019.01.012 (2019).
29. Targher, G., Tilg, H. & Byrne, C. D. Non-alcoholic fatty liver disease: a multisystem disease requiring a multidisciplinary and holistic approach. *Lancet Gastroenterol Hepatol* **6**, 578-588, doi:10.1016/s2468-1253(21)00020-0 (2021).
30. Korten, A.*et al.* Stroke and idiopathic Parkinson's disease: does a shortage of dopamine offer protection against stroke? *Movement disorders : official journal of the Movement Disorder Society* **16**, 119-123, doi:10.1002/1531-8257(200101)16:1<119::aid-mds1024>3.0.co;2-w (2001).
31. Scigliano, G.*et al.* Reduced risk factors for vascular disorders in Parkinson disease patients: a case-control study. *Stroke* **37**, 1184-1188, doi:10.1161/01.STR.0000217384.03237.9c (2006).
32. Gil-Prieto, R.*et al.* Measuring the Burden of Hospitalization in Patients with Parkinson´s Disease in Spain. *PLOS ONE* **11**, e0151563, doi:10.1371/journal.pone.0151563 (2016).
33. Wang, X.*et al.* Comorbidity burden of patients with Parkinson's disease and Parkinsonism between 2003 and 2012: A multicentre, nationwide, retrospective study in China. *Scientific reports* **7**, 1671-1671, doi:10.1038/s41598-017-01795-0 (2017).
34. Chen, H.*et al.* Obesity and the risk of Parkinson's disease. *Am J Epidemiol* **159**, 547-555, doi:10.1093/aje/kwh059 (2004).
35. Dulloo, A. G. & Montani, J. P. Obesity in Parkinson's disease patients on electrotherapy: collateral damage, adiposity rebound or secular trends? *Br J Nutr* **93**, 417-419, doi:10.1079/bjn20041337 (2005).

36. Bu, X.-L.*et al.* The association between infectious burden and Parkinson's disease: A case-control study. *Parkinsonism & Related Disorders* **21**, 877-881, doi:http://dx.doi.org/10.1016/j.parkreldis.2015.05.015 (2015).
37. Fang, F.*et al.* CNS infections, sepsis and risk of Parkinson's disease. *International Journal of Epidemiology* **41**, 1042-1049, doi:10.1093/ije/dys052 (2012).
38. Harris, M. A., Tsui, J. K., Marion, S. A., Shen, H. & Teschke, K. Association of Parkinson's disease with infections and occupational exposure to possible vectors. *Movement Disorders* **27**, 1111-1117, doi:10.1002/mds.25077 (2012).
39. Vlajinac, H.*et al.* Infections as a risk factor for Parkinson's disease: a case–control study. *International Journal of Neuroscience* **123**, 329-332, doi:10.3109/00207454.2012.760560 (2013).
40. Bower, J. H., Maraganore, D. M., Peterson, B. J., Ahlskog, J. E. & Rocca, W. A. Immunologic diseases, anti-inflammatory drugs, and Parkinson disease: a case-control study. *Neurology* **67**, 494-496, doi:10.1212/01.wnl.0000227906.99570.cc (2006).
41. Goldman, S. M.*et al.* Head injury and Parkinson's disease risk in twins. *Ann Neurol* **60**, 65-72, doi:10.1002/ana.20882 (2006).
42. Leopold, N. A., Bara-Jimenez, W. & Hallett, M. Parkinsonism after a wasp sting. *Movement disorders : official journal of the Movement Disorder Society* **14**, 122-127 (1999).
43. Minault, P., Madigand, M. & Sabouraud, O. [Pallidostriatal necrosis after Hymenoptera sting. Parkinsonian syndrome]. *La Nouvelle presse medicale* **10**, 3725-3726 (1981).
44. Boka, G.*et al.* Immunocytochemical analysis of tumor necrosis factor and its receptors in Parkinson's disease. *Neuroscience letters* **172**, 151-154 (1994).
45. Hunot, S.*et al.* FcεRII/CD23 Is Expressed in Parkinson's Disease and Induces, In Vitro, Production of Nitric Oxide and Tumor Necrosis Factor-α in Glial Cells. *The Journal of Neuroscience* **19**, 3440-3447 (1999).
46. McGeer, P. L., Itagaki, S., Boyes, B. E. & McGeer, E. G. Reactive microglia are positive for HLA-DR in the substantia nigra of Parkinson's and Alzheimer's disease brains. *Neurology* **38**, 1285-1291 (1988).
47. Taylor, J. M., Main, B. S. & Crack, P. J. Neuroinflammation and oxidative stress: co-conspirators in the pathology of Parkinson's disease. *Neurochemistry international* **62**, 803-819, doi:10.1016/j.neuint.2012.12.016 (2013).
48. Chen, H.*et al.* Nonsteroidal anti-inflammatory drugs and the risk of Parkinson disease. *Archives of neurology* **60**, 1059-1064, doi:10.1001/archneur.60.8.1059 (2003).
49. Gao, X., Chen, H., Schwarzschild, M. A. & Ascherio, A. Use of ibuprofen and risk of Parkinson disease. *Neurology* **76**, 863-869, doi:10.1212/WNL.0b013e31820f2d79 (2011).
50. Rees, K.*et al.* Non-steroidal anti-inflammatory drugs as disease-modifying agents for Parkinson's disease: evidence from observational studies. *The Cochrane database of systematic reviews*, CD008454, doi:10.1002/14651858.CD008454.pub2 (2011).

51. Wahner, A. D., Bronstein, J. M., Bordelon, Y. M. & Ritz, B. Nonsteroidal anti-inflammatory drugs may protect against Parkinson disease. *Neurology* **69**, 1836-1842, doi:10.1212/01.wnl.0000279519.99344.ad (2007).
52. Jiang, S., Gao, H., Luo, Q., Wang, P. & Yang, X. The correlation of lymphocyte subsets, natural killer cell, and Parkinson's disease: a meta-analysis. *Neurol Sci* **38**, 1373-1380, doi:10.1007/s10072-017-2988-4 (2017).
53. Chen, Y.*et al.* Clinical characteristics and peripheral T cell subsets in Parkinson's disease patients with constipation. *International journal of clinical and experimental pathology* **8**, 2495-2504 (2015).
54. Kustrimovic, N.*et al.* Dopaminergic Receptors on CD4+ T Naive and Memory Lymphocytes Correlate with Motor Impairment in Patients with Parkinson's Disease. *Sci Rep* **6**, 33738, doi:10.1038/srep33738 (2016).
55. Calder, J. S., Holten, I. & McAllister, R. M. Evidence for immune system involvement in reflex sympathetic dystrophy. *J Hand Surg Br* **23**, 147-150, doi:10.1016/s0266-7681(98)80162-9 (1998).
56. Paz, T.*et al.* Hand Function as Predictor of Motor Symptom Severity in Individuals with Parkinson's Disease. *Gerontology* **67**, 160-167, doi:10.1159/000511910 (2021).
57. Paker, N.*et al.* Gait speed and related factors in Parkinson's disease. *J Phys Ther Sci* **27**, 3675-3679, doi:10.1589/jpts.27.3675 (2015).
58. Shulman, L. M.*et al.* The evolution of disability in Parkinson disease. *Movement disorders : official journal of the Movement Disorder Society* **23**, 790-796, doi:10.1002/mds.21879 (2008).
59. Pohar, S. L. & Allyson Jones, C. The burden of Parkinson disease (PD) and concomitant comorbidities. *Arch Gerontol Geriatr* **49**, 317-321, doi:10.1016/j.archger.2008.11.006 (2009).
60. Correia Guedes, L., Mestre, T., Outeiro, T. F. & Ferreira, J. J. Are genetic and idiopathic forms of Parkinson's disease the same disease? *Journal of Neurochemistry* **152**, 515-522, doi:https://doi.org/10.1111/jnc.14902 (2020).
61. Kumar, R. R., Singh, L., Thakur, A., Singh, S. & Kumar, B. Role of Vitamins in Neurodegenerative Diseases: A Review. *CNS Neurol Disord Drug Targets*, doi:10.2174/1871527320666211119122150 (2021).
62. Lima, L. A. R.*et al.* Vitamin D protects dopaminergic neurons against neuroinflammation and oxidative stress in hemiparkinsonian rats. *J Neuroinflammation* **15**, 249, doi:10.1186/s12974-018-1266-6 (2018).
63. Iacopetta, K.*et al.* Are the protective benefits of vitamin D in neurodegenerative disease dependent on route of administration? A systematic review. *Nutr Neurosci* **23**, 251-280, doi:10.1080/1028415x.2018.1493807 (2020).

Figures

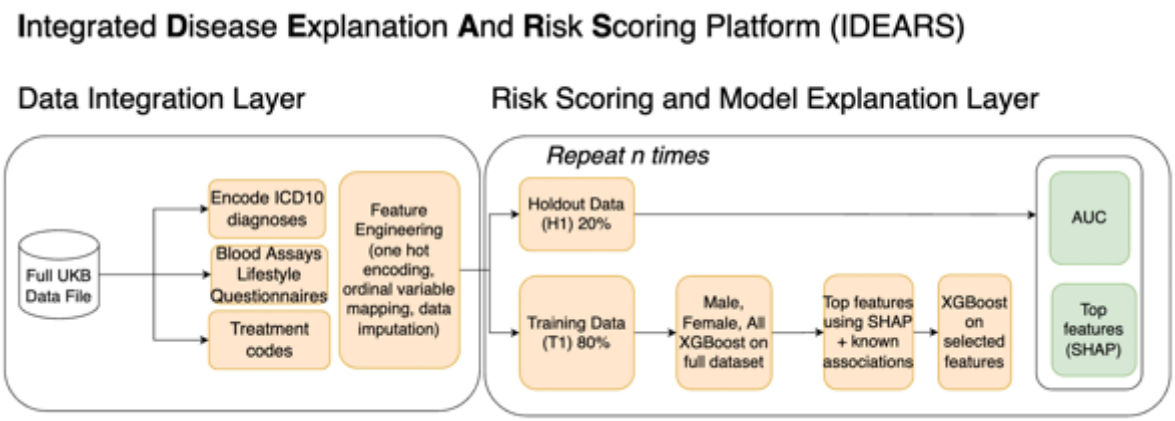

Figure 1

IDEARs platform. An automated platform to facilitate the data integration of large health datasets, performing risk scoring and determining ranked feature importance for any disease.

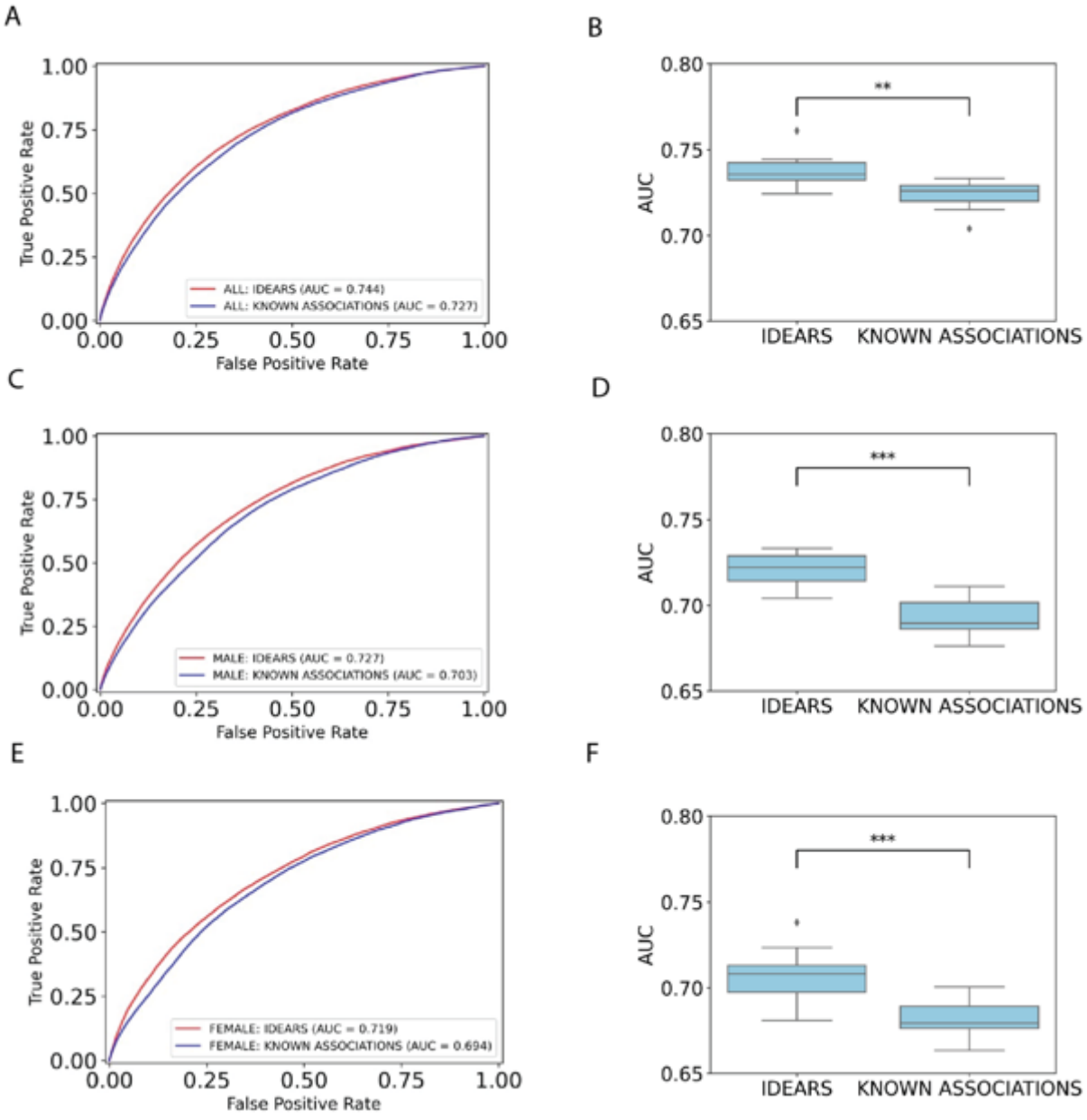

Figure 2

*(**A, C, E**) AUROC graphs for the full IDEARS model, compared to known risk factors based on current meta-analysis[17]. (**B, D, F**) Bar graphs of AUC results from 100 resamples on the combined set of features from the IDEARS model (v1 - set of top 50 SHAP features) compared to all known associations (v2 - which did not appear in the top 50 SHAP features).*

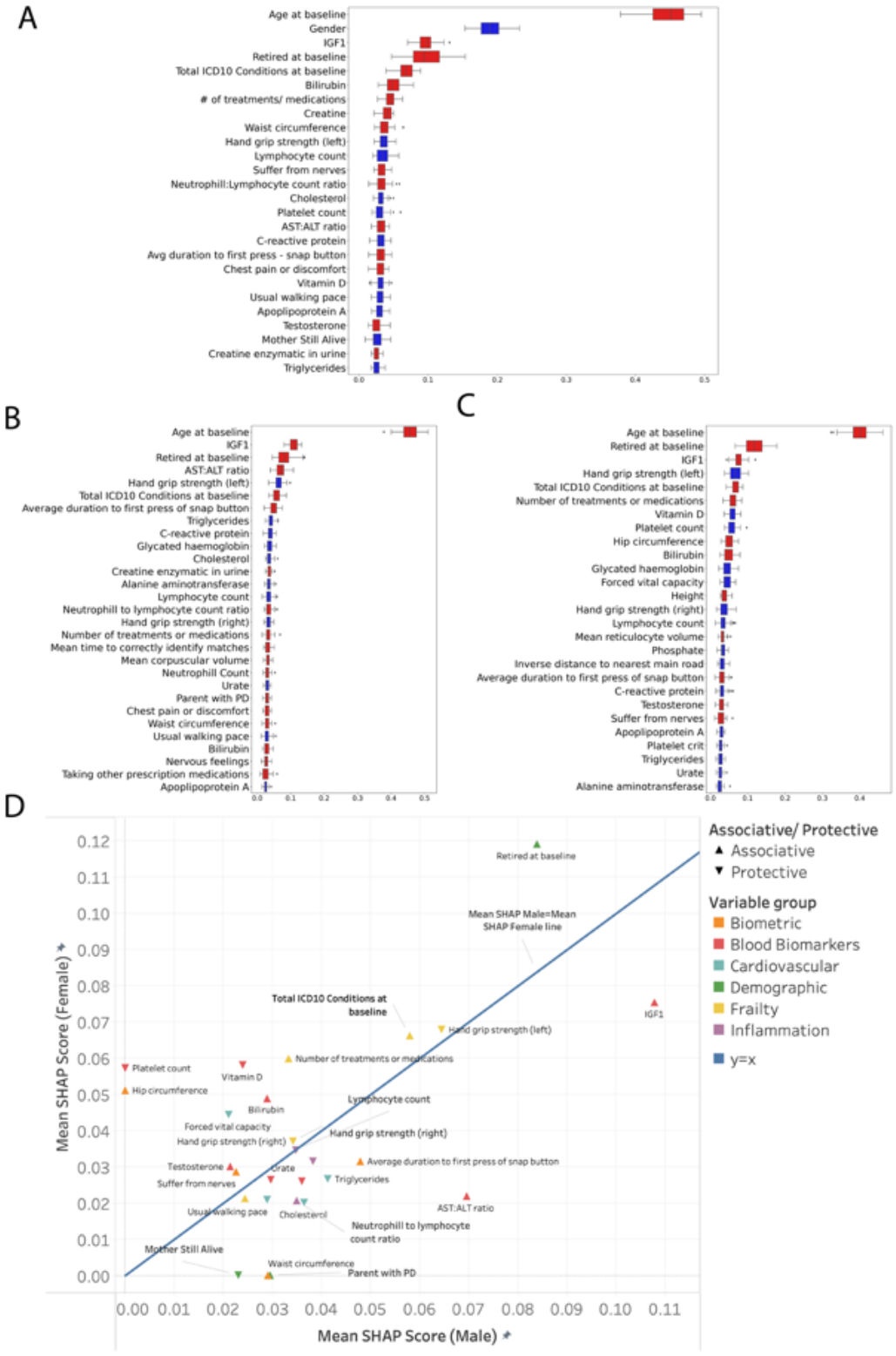

Figure 3

*Box plots indicating the mean SHAP score of the top features from the IDEARS model for the entire cohort (**A**), males only (**B**) and female only (**C**). Those for which a higher value had a positive impact on the model output (i.e. making a PD diagnosis more likely) are coloured red. Those for which a higher value had a negative impact (making PD less likely) are coloured blue. (**D**) Scatter chart with feature importance*

for males versus females NB. age is excluded. Variables are colour coded in groups, and denotes a causative association with PD, denotes a protective association.

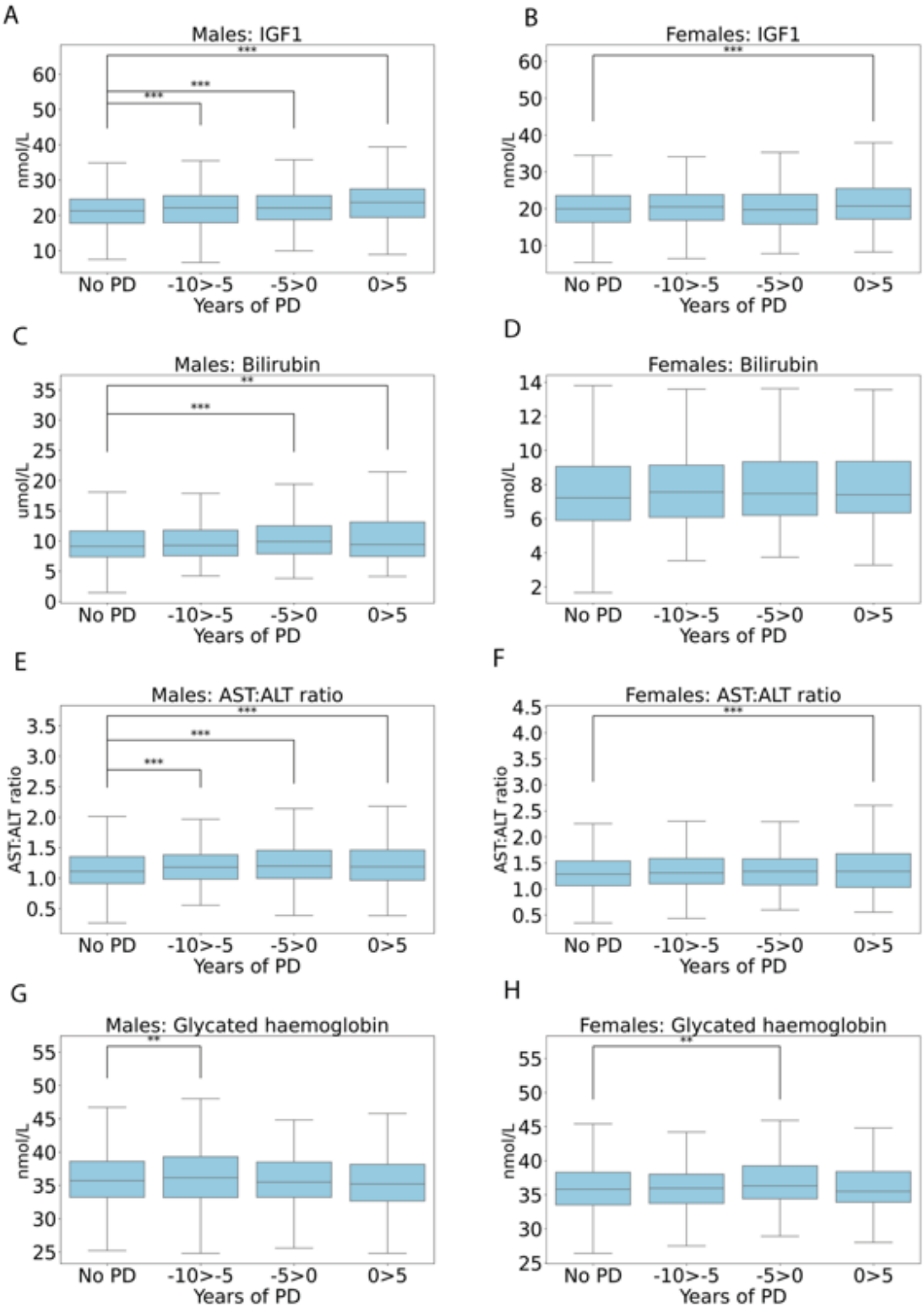

Figure 4

*Box plots showing blood levels of IGF-1 levels and bilirubin, AST:ALT ratio and glycated haemoglobin for (**A, C, E, G**) males and females (**B, D, F, H**) in the 10 years preceding and 5 after a PD diagnosis in 1126*

males and 704 females compared to the non-PD group. AST: aspartate aminotransferase; ALT: Alanine transaminase. Mean +/- SD.

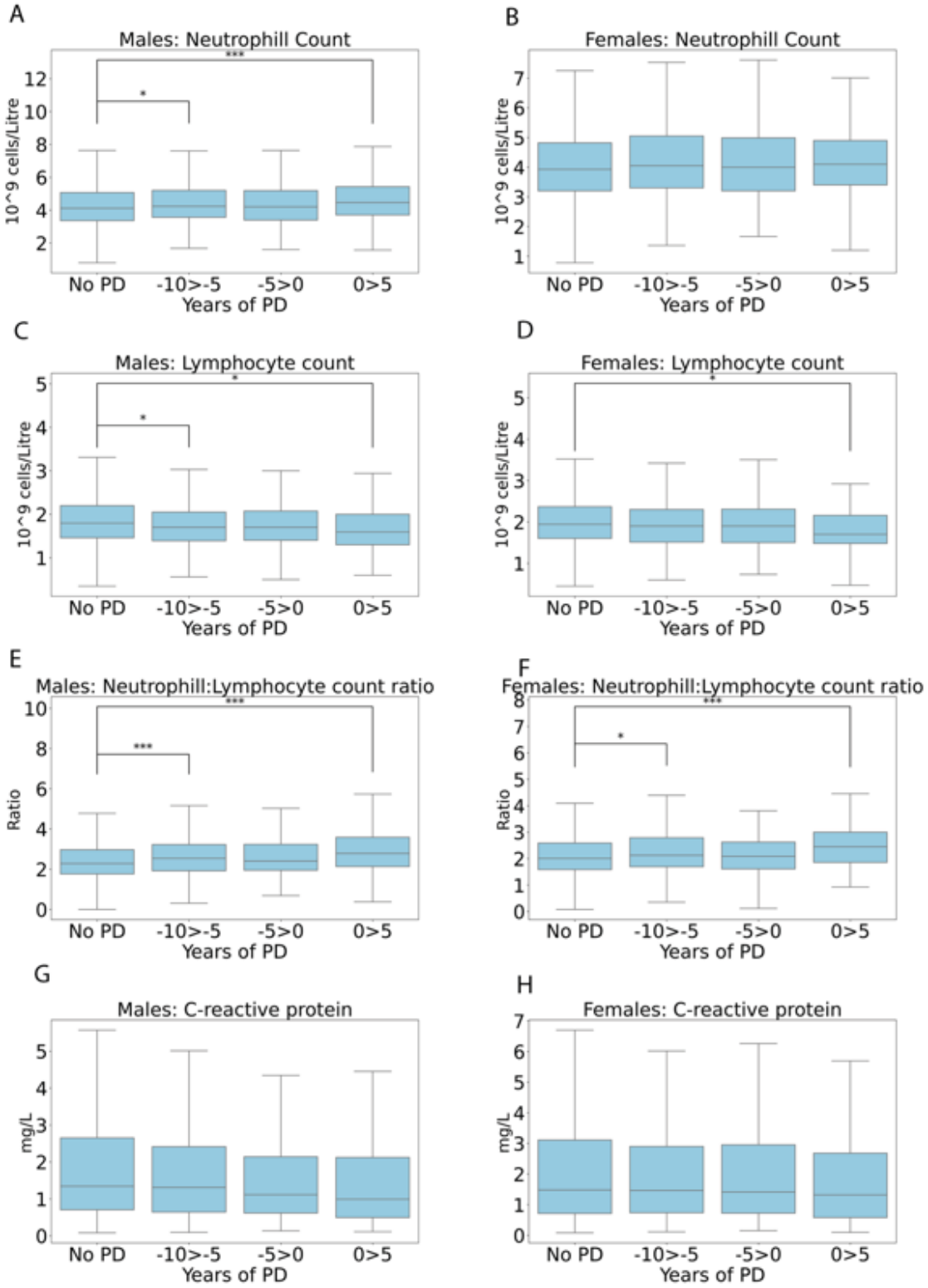

Figure 5

*Box plots showing blood inflammatory markers for (**A, C, E, G**) males and females (**B, D, F, H**) in the 10 years preceding and 5 after a PD diagnosis in 1126 males and 704 females compared to the non-PD group. Mean +/- SD.*

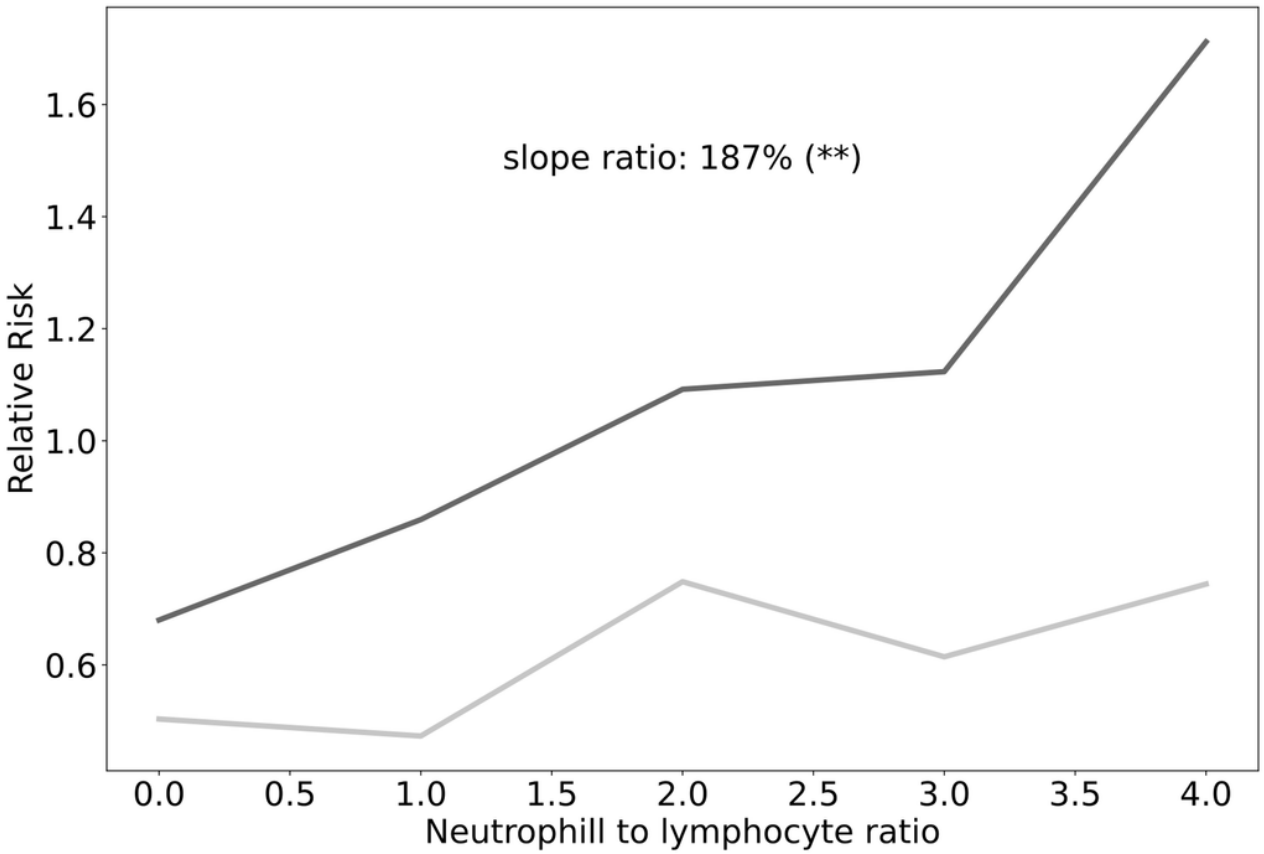

Figure 6

Line graph showing the relationship between the relative risk of being diagnosed with PD and the neutrophil: lymphocyte ratio, between those that were taking the non-steroidal anti-inflammatory, ibuprofen at baseline. Black line - no ibuprofen, grey line – ibuprofen.

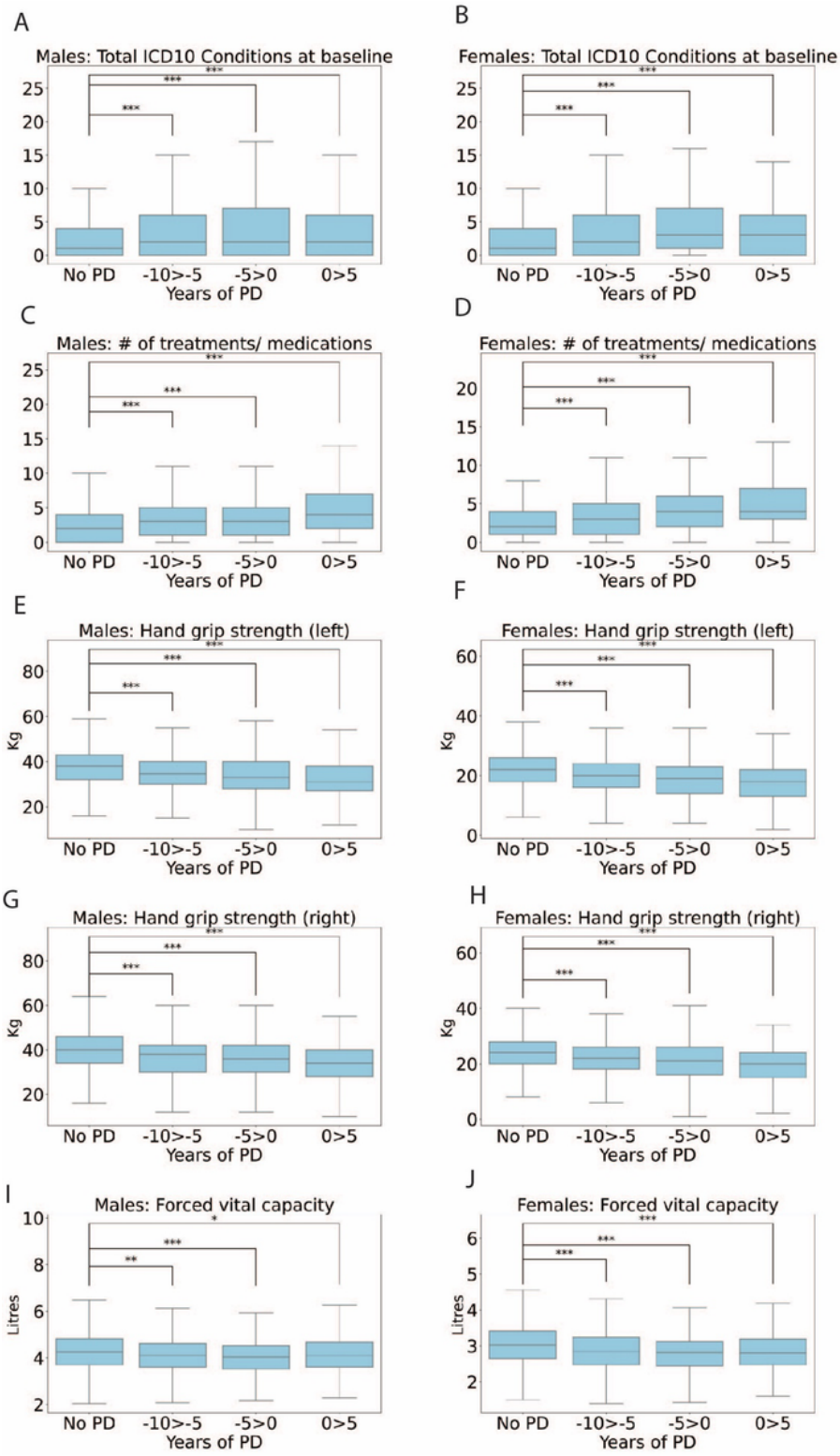

Figure 7

*Box plots showing frailty-related variables for (**A, C, E, G, I**) males and females (**B, D, F, H, J**) in the 10 years preceding and 5 after a PD diagnosis in 1126 males and 704 females compared to the non-PD group. Mean +/- SD.*

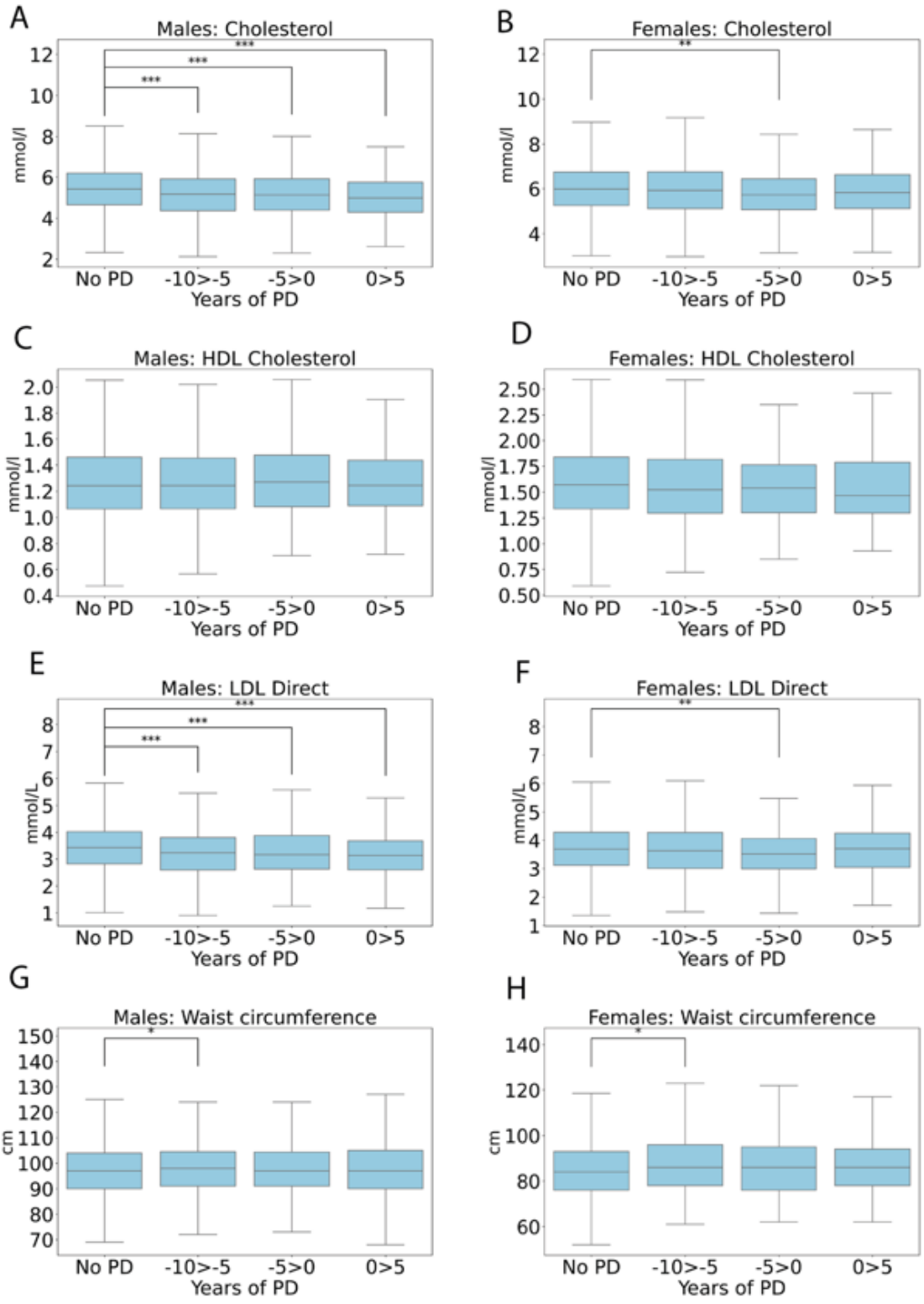

Figure 8

*Box plots showing cardiovascular variables and waist circumference for (**A, C, E, G**) males and females (**B, D, F, H**) in the 10 years preceding and 5 after a PD diagnosis in 1126 males and 704 females compared to the non-PD group. Mean +/- SD.*

Supplementary Files

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

- SupplementarySHAPchartsv2.xlsx
- Supplementaryboxplotsv2.xlsx