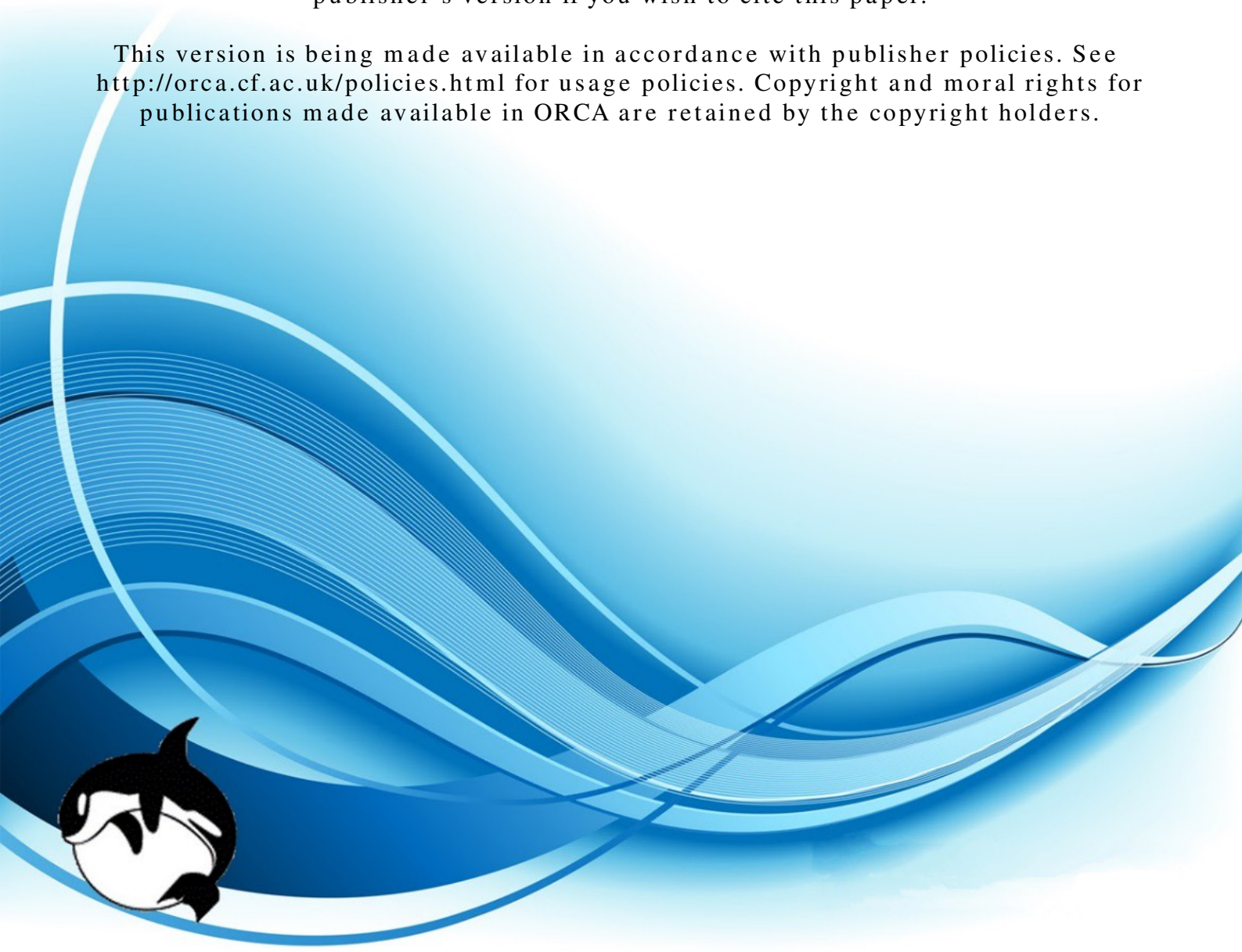

Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis

Ann-Kathrin Schalkamp[1], Kathryn J Peall[2], Neil A Harrison[2,3], Cynthia Sandor[1*]

[1] Psychological Medicine and Clinical Neuroscience, UK Dementia Research Institute, Cardiff University, United Kingdom

[2] Neuroscience and Mental Health Innovation Institute, Division of Psychological Medicine and Clinical Neurosciences, Cardiff, United Kingdom

[3] Cardiff University Brain Research Imaging Centre (CUBRIC), Cardiff, United Kingdom

*Corresponding Author: SandorC@cardiff.ac.uk

Abstract

Parkinson's disease (PD) is a progressive neurodegenerative movement disorder with a latent phase and currently no disease-modifying treatments. Reliable predictive biomarkers that could transform efforts to develop neuroprotective treatments remain to be identified. Using UK Biobank, we investigated the predictive value of accelerometry in identifying prodromal PD in the general population and compared this digital biomarker to models based on genetics, lifestyle, blood biochemistry, and prodromal symptoms data. Machine learning models trained using accelerometry data achieved better test performance in distinguishing both clinically diagnosed PD (N = 153) (area under precision recall curve (AUPRC): 0.14± 0.04) and prodromal PD (N = 113) up to seven years pre-diagnosis (AUPRC: 0.07± 0.03) from the general population (N = 33009) than all other modalities tested (genetics: AUPRC = 0.01± 0.00, p-value = $2.2x10^{-3}$, lifestyle: AUPRC = 0.03± 0.04, p-value = $2.5x10^{-3}$ blood biochemistry: AUPRC = 0.01± 0.00, p-value = $4.1x10^{-3}$, prodromal signs: AUPRC = 0.01± 0.00, p-value =

$3.6x10^{-3}$). Accelerometry is a potentially important, low-cost screening tool for determining people at risk of developing PD and identifying participants for clinical trials of neuroprotective treatments.

Introduction

For most patients diagnosed with Parkinson's disease (PD), 50-70% of nigral dopaminergic neurons will already have degenerated by the time the hallmark motor symptoms manifest and a clinical diagnosis is made [1]. Thus, there remains a need to identify cheap, reliable, easily accessible, and sensitive biomarkers to detect early pathological changes, with success in this field likely to be transformative in identifying suitable participants for clinical trials of neuroprotective therapeutics.

It is well recognised that at the point of a clinical diagnosis, multiple prodromal symptoms could have been present for several years including Rapid-Eye-Movement Sleep Behaviour Disorder (RBD), constipation, hyposmia, depression, anxiety, and excessive daytime somnolence with urinary dysfunction, orthostatic hypotension, sub-threshold motor symptoms, and abnormal dopaminergic molecular brain imaging being more recently added to the criteria for prodromal PD [2-4]. Multiple previous studies have examined these and other markers to determine their sensitivity in identifying prodromal PD [5-7]. However, the absence of multi-modal models, which combine the predictive power of multiple data sources, has limited this work [4]. Furthermore, most studies have tended to compare those with prodromal PD to control cohorts lacking any comorbidity, limiting the translational validity and real-world applicability of these findings. More research is therefore needed to understand the specificity and effective role of prodromal markers in the general population.

Digital sensors passively collect data continuously in real-world settings without added cost or effort [8], and thus obtain robust estimates of a person's impairments and capabilities, and detect subtle changes at the earliest possible opportunity. Such monitoring cannot be achieved through clinical assessments given the limitations of time, cost, accessibility, and sensitivity [9].

Preliminary analyses of acceleration and heart rate data collected with digital sensors have demonstrated the potential for distinguishing those with a clinical diagnosis of PD from those without, as well as the added capabilities of monitoring motor progression, and describing sleeping behaviour [10-12]. However, these quantitative motor measures remain largely understudied [2] with studies often limited by small sample sizes [13], or restricted to analysis only after clinical PD diagnosis [14,15]. Using digital sensors to detect early movement alterations and identify diseases *before* clinical diagnosis is a largely unexplored field with much potential for application in the general population.

This study uses the large, prospective population-based cohort recruited to the UK Biobank (UKBB). Since 2006, data has been collected for >500,000 individuals aged 40-69 years with ongoing passive follow-up of clinical status [16]. Accelerometry data were collected for a randomly chosen subset of this cohort who were approached via email (N = 103,712, collected 2013-2015) [17]. Using these data, we sought to determine whether accelerometry data can serve as a prodromal marker for PD, examining its specificity by comparing data from those receiving a diagnosis of PD or already having a diagnosis of PD to both matched and unmatched unaffected controls as well as individuals diagnosed with related disorders, namely neurodegenerative disorders, movement disorders, and comorbid clinical disorders (Figure 1). We compared the performance of models using accelerometry data to models using data from other modalities, such as genetics, blood, lifestyle, and prodromal symptoms to determine the best data sources to identify prodromal PD in the general population (Extended Figure 1).

Results

UKBB provides an expanding cohort of individuals with PD

Clinical diagnoses within the UKBB currently derive from multiple sources including self-reported symptoms and diagnoses, hospital records, death records, and primary care data. However, there is no clinical diagnostic validation and the data has incomplete coverage. To

ensure that the PD cohort identified was neither an over- or under-estimation, we compared the prevalence in our cohort to that expected based on 2015 UK population statistics [18].

A total of ~0.76% of UKBB participants had been diagnosed with PD by March 2021. Overall, the observed number of PD cases was slightly lower than that expected based on 2015 UK population statistics (2015 expected: 2252.61, 2015 observed: 1984; Extended Figure 2 and Figure 2), with 5255 expected by 2030 assuming no further deaths were to occur. In general, we found that the prevalence and incidence of PD in the UKBB cohort closely resembled that expected from such a population.

Acceleration is reduced several years prior to diagnosis

We compared the average acceleration for each hour of the day over a 7-day period between the diagnostic groups. At the time of or within two years after accelerometry data collection, 273 participants were diagnosed with PD (mean years since diagnosis: 5.04±6.37; Supplementary Table 1, Figure 3A). An additional 196 individuals received a new PD diagnosis more than two years after accelerometery data collection (mean years to diagnosis: 4.33±1.30, Figure 3A). The prodromal group was significantly older than the diagnosed group (t-statistic = 3.26, d.o.f. = 453, p-value = 1.2×10^{-3}, 95% CI = [0.65, 2.61], Cohen's d = 0.298) and were therefore not directly compared in this initial analysis. We randomly sampled age- and sex-matched unaffected controls (1:1) for each participant diagnosed with PD and each participant that would go on to get a diagnosis of PD. Prodromal (from 7am: p-value = 1.9×10^{-4} to 12am: p-value = 4.5×10^{-4}) and diagnosed PD cases (from 7am: p-value = 3.7×10^{-5} to 12am: p-value = 1.4×10^{-3}) both showed a significantly reduced acceleration profile over all hours between 7am and 12am than their age- and sex-matched unaffected controls (Figure 3B, Supplementary Table 3). No differences in average acceleration during the hours from 12am to 7am were found (Supplementary Table 3). A reduction in acceleration during daytime could thus be observed several years prior to clinical diagnosis.

No other disorder shows a similar reduction before diagnosis

As physical activity varies between individuals irrespective of health status, we explored whether the observed reduction in acceleration was unique to PD or whether it could also be observed in other clinical disorders, notably other neurodegenerative and/or movement disorders (Supplementary Table 1). We calculated residual average acceleration corrected for age, sex, and BMI via linear regression models trained on unaffected controls (N = 36058) (Supplementary Table 4). As anticipated, several participants were diagnosed with multiple comorbidities, hence those with a comorbid PD diagnosis or a co-diagnosis of depression were removed.

There was a significant reduction in residual average acceleration in diagnosed PD (t-statistic = 6.25, p-value = 4.17×10^{-10}) and prodromal PD (t-statistic = 5.69, p-value = 1.3×10^{-8}) compared to unaffected controls (Figure 3C, Supplementary Table 10), but no significant difference between individuals with prodromal and diagnosed PD (p-value = 0.88). We did not find a significant effect of PD treatment on average acceleration in those diagnosed with PD (Supplementary Figure 4). Of those investigated, 'Depression' was the only other disorder found to show a reduction in acceleration following diagnosis. None of the disorders investigated were found to have a reduction in acceleration prior to diagnosis, as was observed for PD (Figure 3C, Supplementary Table 10). Overall, the finding of a reduction in acceleration both prior to and following diagnosis was unique to PD, suggesting this measure to be disease specific with potential for use in early identification of individuals likely to be diagnosed with PD.

Sleep is more disrupted in PD than in other disorders

We downloaded and extracted sleep features from raw accelerometry data for individuals with any of the discussed disorders and unaffected controls; data from a total of 65901 individuals were processed. We labelled physical activity classes using a pretrained Random Forest [19] and derived features for each activity class with night-time 11pm to 6:59am and daytime 07am

to 10:59pm. We corrected these extracted features for age, sex, and BMI as learned from the unaffected controls (N = 36058) with linear regression models. Sleep features derived from acceleration data indicated reduced quality and duration of sleep both in the prodromal phase and after having been diagnosed with PD (Figure 4, Supplementary Tables 11-15). Compared to both unaffected controls and the prodromal PD group, individuals with a diagnosis of PD slept for fewer hours overall (controls: p-value = 1.59×10^{-10}, prodromal: p-value = 1.78×10^{-5}), had fewer consecutive hours of sleep (controls: p-value = 9.66×10^{-38}, prodromal: p-value = 2.98×10^{-5}), and slept more frequently during the day (controls: p-value = 1.02×10^{-30}, prodromal: p-value = 3.13×10^{-5}). Prodromal PD cases (p-value = 5.62×10^{-7}) and those diagnosed with PD (p-value = 6.5×10^{-4}) woke up more frequently during the night compared to unaffected controls, with no significant difference between prodromal and diagnosed PD cohorts (p-value = 0.22). Individuals with prodromal PD slept longer than unaffected controls (p-value = 1.59×10^{-10}) and diagnosed PD cases (p-value = 1.78×10^{-5}).

Examination of the other diagnostic cohorts identified less deterioration in the sleep measures than were found in PD (Figure 4, Supplementary Tables 11-15). Across the distinct diagnostic groups, the length of uninterrupted sleep was the most frequently observed feature to differ between groups. The number of nocturnal awakenings were higher in those in the 'AllCauseDementia', 'Osteoarthritis', and 'Depression' cohorts following diagnosis however, none of the diagnostic groups examined presented a reduction in this feature at the prodromal stage, as was found for PD.

Accelerometry data predicts prodromal PD

We next explored the predictive power of accelerometry data following the TRIPOD reporting guidelines [20] in terms of area under precision recall curve (AUPRC) at an individual level using Lasso logistic regression models with average acceleration, age, and sex as features (Supplementary Table 5) in three control group settings: matched unaffected controls, all unaffected controls (N = 24987), or the general population (N = 33009). The AUPRC was

chosen due to the unbalanced datasets [21]. The class prevalence was denoted as N_cases/(N_cases+N_control). Average acceleration distinguished those diagnosed with PD (N = 153) from matched unaffected controls (N = 153) with a mean AUPRC of 0.78±0.06 and could distinguish prodromal PD (N = 113) from matched unaffected controls (N = 113) with the same performance. Identifying diagnosed PD and prodromal PD from non-matched unaffected controls (N = 24987) led to performances of 0.09±0.05 (prevalence = 0.0061) and 0.09±0.02 (prevalence = 0.0045) respectively. Diagnosed PD and prodromal PD could also be identified from a general population cohort including all unaffected controls, and prodromal and diagnosed cases of 'Osteoarthritis', 'Dystonia', 'OtherParkinsonism', and 'AllCauseDementia' (N = 33009) using only average acceleration with AUPRCs of 0.05±0.04 (prevalence = 0.0034) and 0.06±0.05 (prevalence = 0.0046) respectively. Adding derived physical activity and sleep features (Supplementary Table 5) increased performance of the models to identify individuals diagnosed with PD from the general population to 0.14±0.04 AUPRC and to 0.07±0.03 to identify prodromal PD from the general population. The increase in performance compared to models using average acceleration was only significant for the diagnosed PD model (p-value = 0.01). The most robustly selected feature in all settings was mean acceleration during epochs classified as light physical activity which reduced the risk of having/getting PD (Extended Figure 3E-5E, Supplementary Figure 13E-15E, Extended Figure 6); meaning that slowness of movement during normal physical activity is predictive for both, prodromal and diagnosed PD.

Accelerometry data outperforms known PD prodromal markers

Several modalities have been explored previously for their value in identifying prodromal PD; however, these were often investigated in isolation and in clinically refined cohorts, rather than the general population. Here, we examined genetics, lifestyle, blood biochemistry, recognised prodromal symptoms for PD, as well as accelerometry (Table 1). We trained Lasso logistic regression models on these different modalities to identify diagnosed (N = 153) or prodromal (N = 113) PD from the three different control groups (Supplementary Table 9). Models trained on lifestyle, serum biochemical blood markers, recognised prodromal symptoms, or genetic

factors showed lower AUPRC scores than models trained on accelerometry features, potentially due to its higher disease specificity (Figure 5A-C, Supplementary Table 5, Extended Figure 7). When comparing each modality-specific model to the accelerometry modality, significant improvement was found compared to genetics, lifestyle, and blood biochemistry in all settings (Supplementary Table 6). For example, for identifying prodromal cases from the general population: genetics achieved an AUPRC of 0.01± 0.00 (p-value = $2.2x10^{-3}$), lifestyle an AUPRC of 0.03± 0.04 (p-value = $2.5x10^{-3}$), and blood biochemistry an AUPRC of 0.01± 0.00 (p-value = $4.1x10^{-3}$). Significant improvement compared to prodromal signs only became apparent in the general population control setting, where prodromal signs achieved an AUPRC of 0.01± 0.00 (p-value = $3.6x10^{-3}$).

We further compared each modality-specific model to a no-skill classifier (predictors: intercept) (Supplementary Figure 5 & 6) and noted genetics, lifestyle, and blood biochemistry in some settings not outperforming this baseline (Supplementary Table 6). We next compared each single modality model to its respective combined one where the accelerometry modality was added to the predictors. In the diagnosed PD models, the combined models always outperformed the single models for all control settings (Supplementary Figure 5). For prodromal PD, adding accelerometry to prodromal symptoms only led to an improvement in the general population setting (Supplementary Figure 6). The prodromal symptoms modality hence showed similar performance in identifying prodromal PD when the controls did not include participants diagnosed with related disorders. A model combining all available modalities did not outperform the single accelerometry model performance in any setting, which could be due to each modality capturing different degrees of the same information (Supplementary Figure 17, 18). Overall, the accelerometry modality performed best, especially in the general population setting.

We next evaluated which factors within each modality were considered the most relevant (Extended Figure 3-5, Supplementary Figure 13-15) and provided a measure of how many

features from within a modality were significant in the combined model. We restricted this second analysis to the models using matched unaffected controls. The most important features in the combined model resembled those found in the modality-specific models (Extended Figure 3-5, Supplementary Figure 13-15). Features from the accelerometry modality made up the largest portion (67% for the model identifying prodromal PD and 50% for the model identifying diagnosed PD) of the most important features in the combined model (Extended Figure 3F, Supplementary Figure 13F). The second modality, in order of importance to the model was genetic markers with the PRS for PD as the important feature. As the accelerometry modality has the highest number of features compared to the other modalities (Table 1), one could argue that their importance was purely due to their dimensionality. We investigated this through a stacked model where the predicted probabilities of the modality-specific models served as input to a final logistic regression model. This also identified the accelerometry as the most important one (Supplementary Figure 16).

Accelerometry data predicts time to PD diagnosis

An estimate of time to diagnosis would not only have potential clinical utility but would also be important in clinical trials evaluating the efficacy of disease-modifying or curative therapies. As such, we next explored which modality would be most beneficial in predicting time to clinical diagnosis of PD. A survival random forest model was used to predict time to diagnosis using the same modality-specific modelling approach described previously for the logistic regression models, here, however, we only focus on prodromal PD and modelling their time to diagnosis compared to our three control groups (Extended Figure 8). We chose the time-dependent AUROC to evaluate our models, and provided a measure of how well a model can identify all pheno-converted cases versus controls up to specified time points. The model trained on accelerometry features achieved a mean AUROC of 0.74 ± 0.04 when restricted to matched unaffected healthy controls, 0.86 ± 0.06 when prodromal cases are identified from all unaffected controls, and 0.84 ± 0.04 when trained on the general population (Figure 5D-F, Supplementary Table 7). The brier score can be found in Supplementary Figure 20. As survival class (right

censored class) outnumbered the number of people receiving a diagnosis of PD within the observed time frame, the AUROC scores should only be evaluated in a comparative manner as this metric can be overoptimistic in this imbalanced setting. The accelerometry model could predict the probability of not receiving a PD diagnosis over time significantly better than any other single modality model in all settings and performed similarly to the combined model (Figure 5D-F, Supplementary Table 8). This finding highlights that acceleration data not only allowed us to predict who would develop PD but also when this diagnosis might be expected.

Discussion

Here we show the potential of accelerometry as a biomarker to screen for PD. We found that reduced acceleration manifests years prior to clinical PD diagnosis. This pre-diagnosis reduction in acceleration was unique to PD and was not observed for any other disorder examined. By comparing the predictive value of accelerometry with other modalities including genetics, lifestyle, blood biochemistry, and prodromal symptoms, we found that no other data modality performed better in identifying future diagnosis of PD. This improvement in predictive power was most clearly visible when assessing our models in a real-world scenario where the control group contained individuals with related disorders. Finally, we showed that accelerometry can predict the time-point at which a PD diagnosis can be expected.

This work builds on prior clinical data which has demonstrated abnormal motor functioning during the prodromal phase of PD. Darweesh, et al. [22] showed that impairment in activities of daily living and signs of slowness appeared up to seven years prior to a clinical diagnosis. Similarly, Fereshtehnejad, et al. [7] analysed the temporal evolution of multiple prodromal markers in a longitudinal cohort of patients diagnosed with RBD and highlighted the predictive potential of early motor symptoms to identify prodromal PD up to six years prior to diagnosis. However, these works were limited to either a high risk RBD population or assessments made in the clinical setting, which require the increased cost and time of an in-person visit.

Previous work has already explored the use of digital gait markers for diagnosing PD [13,14,23]. For example, also using UKBB data Williamson, et al. [14] demonstrated high accuracy in detecting diagnosed PD cases. However, they focused on prevalent PD cases and did not explore the possibility of using these digital markers to identify PD *before* clinical diagnosis. To date, the only other study to have used gait-measuring sensors to investigate the prodromal PD was limited to 16 subjects [13]. Further, the gait data acquired in that study was collected in clinic during specified tasks.

Other digital markers have been investigated for their use as potential prodromal biomarkers. For example, nocturnal breathing patterns that can be collected at home via radio waves have been shown to identify 75% of the 12 prodromal cases as PD before their clinical diagnosis [23]. Cognitive and functional impairment prior to diagnosis have been assessed in several neurodegenerative disorders, including PD, using UKBB data [24]. Transferability to the general population and disease specificity however has not been reported in previous studies using accelerometry data.

Overall, we identified five major gaps in research that our current work aimed to address: studying the (1) prodromal phase of PD using passively collected (2) real-world gait-sensor-based data in a (3) large sample size (4) while comparing its performance to other established markers and (5) its generalizability to the general population. To our knowledge, by using a large sample of individuals who convert to PD *after* data collection, we provided the first demonstration of the clinical value of accelerometry-based biomarkers compared to other markers for prodromal PD in the general population.

Using accelerometry data for screening in the general population is feasible as the data is easily accessible. Smart-devices capable of collecting accelerometry data are used daily by most people [25].Challenges to overcome include measurement validity and capability, data privacy, and liability concerns [26].Further, processing of the vast amounts of data generated by

digital sensors is resource and time intense. As we have demonstrated here, a single week period of data is predictive for several future years, and therefore longer intervals between assessment could be employed, reducing resource demands. If these limitations are addressed, wearables and other health-sensor devices hold the ability to transition medicine into a digital health era improving accessibility in remote areas, reducing cost, and improving healthcare [27].

There are several limitations in this study, the primary one being the lack of external replication, although extensive cross-validation was performed to attempt to mitigate against any cohort specific biases. This largely related to the lack of another dataset equivalent to UKBB in terms of scale and volume of data that would allow the prodromal phase of multiple disorders to be retrospectively studied. For example, though the Parkinson's Progression Marker Initiative (PPMI)[28] cohort provides longitudinal smart watch data for a prodromal cohort of 158 cases, here prodromal is defined as people at risk, not people who subsequently received a PD diagnosis. Another dataset, 'The All of Us' cohort[29] could provide a valuable future replication resource, however is currently limited by the small number of people who have received a diagnosis after data collection. This is due, in part, to recruitment being from a wider age range (>18 years in 'The All of Us' compared to 40-69 years in the UKBB) with data capture to date being over a shorter period (2018 onwards for 'The All of Us' study and 2006 onwards for the UKBB).

Several restraints concerning data availability within the UKBB should be noted. For the majority, accelerometry data was only collected over one seven-day period. Longitudinal data on acceleration would allow investigation of individual trajectories. Additionally, several clinically recognised prodromal markers, such as dopamine transporter imaging or motor examinations, were not available within the UKBB and therefore, could not be compared to the accelerometry data, despite their recognised high predictive power [2,30]. As we chose the time of accelerometry data collection as the defining time point for the group assignment, this limits

the comparison to other data, like lifestyle and blood biochemistry, collected prior to this. Further, not all included features were available for all participants and hence, models were trained only on a subset of individuals where complete information was available, artificially reducing our sample size but allowing greater comparability between models (Table 1). Of note, data availability in the UKBB does not reflect the real-world availability of the modalities. For example, genetic data is more sparsely available in real-life but was prioritised within the UKBB, while accelerometry data is gathered for many people in real-life but only for a subset in UKBB.

Downloading and processing of the raw accelerometry data is time-consuming (30 seconds to download data from one participant (~250MB), ~3 minutes to process), thus we restricted our analysis to the identified diagnoses of interest and unaffected controls. This limits the transferability of our model to clinical practice as it was not trained on a perfect representation of the general population. A final limitation is our choice of model. Using Lasso logistic regression, we focussed on an interpretable model of low complexity. Using class weighting, we prioritised sensitivity over specificity in our model training, thus creating a screening tool rather than a replacement for clinical diagnosis (Extended Figure 9). Exploring more advanced models that allow for non-linearity could potentially further increase the performance of the models.

In conclusion, our results suggest that accelerometry collected with wearable devices in the general population could be used to identify those at elevated risk for PD on an unprecedented scale, and, importantly, these individuals who will likely convert within the next few years can be included in studies for neuroprotective treatments.

Acknowledgements

We are grateful for the Advanced Research Computing at Cardiff. We are also grateful for the valuable comments of Caleb Webber and the input on survival modelling from Valentina

Escott-Price. **A.-K. Schalkamp** is supported by a PhD studentship funded by the Welsh Government through Health and Care Research Wales (HS-20-11)

C. Sandor is supported by the UK Dementia Research Institute (UK DRI) funded by the Medical Research Council (MRC), Alzheimer's Society and Alzheimer's Research UK (AR-UK) and by the Ser Cymru II programme (CU187) which is part-funded by Cardiff University and the European Regional Development Fund through the Welsh Government.

K. Peall is funded by an MRC Clinician-Scientist Fellowship (MR/P008593/1).

N. Harrison has nothing to declare.

- HS-20-11 (A.-K.S.)
- MR/P008593/1 (K.P.)
- CU187 (C.S.)

Author Contribution

A-K.S. and C.S. participated in designing the study, topic definition, and review of relevant studies. Machine learning models and statistical analyses were designed and implemented by A-K.S.. Figures and tables were done by A-K.S. with the support of C.S.. A-K.S. wrote the first draft. A-K.S., C.S., N.A.H., and K.J.P. contributed to subsequent versions of the manuscript. All authors critically reviewed the paper, all authors have a clear understanding of the content, results, and conclusions of the study and agree to submit this manuscript for publication. The corresponding author (C.S.) declares that all authors listed meet the authorship criteria and that no other authors involved in this study are omitted. C.S. is ultimately responsible for this article.

Competing interests

The authors declare no competing interests.

Tables

modality	features	N features

No-skill	Intercept	1
Genetics & Family history	Polygenic risk scores of 34 traits Family history of: Stroke, Diabetes, Severe Depression, Alzheimer's disease Dementia & Parkinson's disease	42
Blood biochemistry	Albumin, Alkalinephosphatase, Alanineaminotransferase, ApolipoproteinA, ApolipoproteinB, Aspartateaminotransferase, Urea, Calcium, Cholesterol, Creatinine, C reactiveprotein, CystatinC, Gammaglutamyltransferase, Glucose, Glycatedhaemoglobin HbA1c, HDLcholesterol, IGF 1, LDLdirect, Phosphate, SHBG, Totalbilirubin, Testosterone, Totalprotein, Triglycerides, Urate, Vitamin D	29
Lifestyle	AlcoholStatus Current, AlcoholStatus Previous, SmokeStatus Current, SmokeStatus Previous, Daytime Sleepiness Often, AlcoholFrequency LessThanWeekly, BMI, Waist Circumference, Hip Circumference, Diastolic BloodPressure, PulseRate, Body-Fat Percentage, TownsendDeprivationIndex	16
Prodromal Symptoms	UrinaryIncontinence, Constipation, ErectileDysfunction, Anxiety, REM Behavioral Sleep Disorder (RBD), Hyposmia, OrthostaticHypotension, Depression	11
All Accelerometry features	UKBB provided averages, weartime-bias corrected value, and standard deviations for days and hours, self-derived features for physical activity epochs (sleep, sedentary, light, MVPA, imputed)	82
Combined	union of above	168
Stacked	predicted probabilities from single-modality models (genetics + family, blood, lifestyle, prodromal symptoms, all accelerometry, intercept)	6

Table 1: Feature set for each of the modalities.

For each modality we show which and how many predictors were included. Every modality includes the covariates (age at accelerometry data collection, sex) and an intercept.

Figure Legends

Figure 1: Overview of performed analyses – subject flowshart

We show for each analysis how many subjects were included and why others were removed. Starting with the complete UK Biobank dataset, we first focus on those with accelerometry data available. Those are assigned to groups based on our diagnosis extraction method. Three different analyses follow. The first one being done on raw data where unaffected controls are matched to each prodromal and diagnosed case. The second one encompasses statistical analyses for group comparisons, where first only subjects with information on covariates were kept such that residuals could be computed, and then cases with comorbid depression and Parkinson's disease were removed. The third analysis trains the prediction models where only subjects with complete information on all predictors were kept.

Figure 2: Estimated and observed prevalence of Parkinson's disease in UK Biobank.

Estimated (dashed) and observed (solid) number of people living with Parkinson's disease in the UK Biobank over time within age groups is shown. Estimated number of cases uses the population-based UK statistics from 2015 [18].

Figure 3: Reduction in acceleration prior to diagnosis is unique to Parkinson's disease

[A] Baseline data were collected between 2006 and 2010; accelerometry data was gathered for a subset between 2013 and 2015. Diagnosed cases (green) were diagnosed prior to or within the subsequent two years of accelerometry data collection. Prodromal cases (orange) were diagnosed two or more years after accelerometry data collection. [B] Average acceleration in milligal (0.01 mm/s^2) is shown in one-hour intervals over the course of one day. Group means for prodromal participants (N = 196, orange, dashed), unaffected controls matched to the prodromal ones (N = 196, blue, dashed), diagnosed participants (N = 273, green, solid), and unaffected controls matched to the diagnosed ones (N = 273, blue, solid) is plotted with the respective 95% confidence interval. [C] Boxplots for residual (age-, BMI-, and sex-corrected through unaffected control cohort) no-wear time bias corrected average acceleration after removal of cases diagnosed with comorbid depression or PD are shown for seven disease groups and unaffected controls. Analyses without these adjustments can be found in Supplementary Figure 2. For each disease group we differentiate between diagnosed (green), prodromal (orange), and healthy

(blue). The number of individuals in each group is indicated in the central box. The boxplots show the mean as the centre, the 25% and 75% quartiles as the bounds of the box, and the Q3 + 1.5*IQR/ Q1 - 1.5*IQR as the whiskers. P-values are shown for group differences (two sided Welch T-test) where significance is reached with a 0.05 Bonferroni-corrected threshold of 2.38×10^{-3}.

Figure 4: Quality and duration of sleep are reduced in diagnosed but not prodromal Parkinson's or any other disorder

Boxplots for five residual (age-, BMI-, and sex-corrected through unaffected control cohort) accelerometry derived sleep features after removal of cases diagnosed with comorbid depression or PD are shown for five disease groups and unaffected controls displaying the mean as the centre, the 25% and 75% quartiles as the bounds of the box, and the Q3 + 1.5*IQR/ Q1 - 1.5*IQR as the whiskers. Supplementary Figure 3 shows the same analyses without exclusion of cases diagnosed with comorbid depression. Due to the covariate correction including a subtraction of the effects of the covariates the residual is displayed and does not reflect the true value of the variable leading to potentially negative values. For each disease group we differentiate between diagnosed (green), prodromal (orange), and healthy (blue). The number of individuals in each group is indicated in the central box. P-values are shown for group differences (two sided T-test, two sided Welch T-test for comparisons including Healthy group) where significance is reached with a 0.05 Bonferroni-corrected threshold of 2.38×10^{-3}.

Figure 5: Accelerometry identifies Parkinson's disease and predicts time to diagnosis better than any other risk factors.

[A-C] Bar plots indicate the performance of each logistic regression model using different feature sets (Table 1). The mean area under precision recall curve (AUPRC) across the five outer cross-validation folds is plotted with the error bars indicating the Bonferroni-adjusted 95% confidence interval. The individual performances per fold are shown via dots overlayed on the bars. We show this for a no-skill, five single modality models, and one combined model for three different tasks with three different control groups, [A] matched unaffected controls, [B] all unaffected controls, [C] general population. Significance of group differences (two sided T-test (N=5)) for each model compared to the all accelerometry model are indicated with star symbols, where significance is reached with a 0.05 Bonferroni-corrected threshold of 8.33×10^{-3} (ns: $8.33 \times 10^{-3} < p <= 1$, *: $8.33 \times 10^{-4} < p <= 8.33 \times 10^{-3}$, **: $8.33 \times 10^{-5} < p <= 8.33 \times 10^{-4}$, ***: $8.33 \times 10^{-6} < p <= 8.33 \times 10^{-5}$, ****: $p <= 8.33 \times 10^{-6}$). [D-F] A performance

evaluation of the survival models is provided in a time-dependent manner, where the performance is assessed in identifying all pheno-converted cases versus not yet converted or censored controls up to specified time points. The mean time-dependent area under the receiver operator curve (AUROC) of the random survival forests is plotted for several evaluation time-points (years since data collection) together with Bonferroni-adjusted 95% confidence interval derived from the five outer cross validation folds for seven years since accelerometry data collection. We show this for [D] a control group made up of matched unaffected controls, [E] a control group including all unaffected controls, and [F] a control group representing the general population.

References

1. Fearnley, J.M. & Lees, A.J. Ageing and Parkinson's disease: substantia nigra regional selectivity. *Brain* **114 (Pt 5)**, 2283-2301 (1991).
2. Heinzel, S.*, et al.* Update of the MDS research criteria for prodromal Parkinson's disease. *Mov Disord* **34**, 1464-1470 (2019).
3. Postuma, R.B. & Berg, D. Advances in markers of prodromal Parkinson disease. *Nat Rev Neurol* **12**, 622-634 (2016).
4. Postuma, R.B. & Berg, D. Prodromal Parkinson's Disease: The Decade Past, the Decade to Come. *Mov Disord* **34**, 665-675 (2019).
5. Hustad, E. & Aasly, J.O. Clinical and Imaging Markers of Prodromal Parkinson's Disease. *Front Neurol* **11**, 395 (2020).
6. Fereshtehnejad, S.M.*, et al.* Validation of the MDS research criteria for prodromal Parkinson's disease: Longitudinal assessment in a REM sleep behavior disorder (RBD) cohort. *Mov Disord* **32**, 865-873 (2017).
7. Fereshtehnejad, S.M.*, et al.* Evolution of prodromal Parkinson's disease and dementia with Lewy bodies: a prospective study. *Brain* **142**, 2051-2067 (2019).
8. Brognara, L., Palumbo, P., Grimm, B. & Palmerini, L. Assessing Gait in Parkinson's Disease Using Wearable Motion Sensors: A Systematic Review. *Diseases* **7**(2019).
9. Dorsey, E.R.*, et al.* Deep Phenotyping of Parkinson's Disease. *J Parkinsons Dis* **10**, 855-873 (2020).
10. Shah, V.V.*, et al.* Digital Biomarkers of Mobility in Parkinson's Disease During Daily Living. *J Parkinsons Dis* **10**, 1099-1111 (2020).
11. Schlachetzki, J.C.M.*, et al.* Wearable sensors objectively measure gait parameters in Parkinson's disease. *PLoS One* **12**, e0183989 (2017).
12. Johansson, D., Malmgren, K. & Alt Murphy, M. Wearable sensors for clinical applications in epilepsy, Parkinson's disease, and stroke: a mixed-methods systematic review. *J Neurol* **265**, 1740-1752 (2018).
13. Del Din, S.*, et al.* Gait analysis with wearables predicts conversion to parkinson disease. *Ann Neurol* **86**, 357-367 (2019).
14. Williamson, J.R., Telfer, B., Mullany, R. & Friedl, K.E. Detecting Parkinson's Disease from Wrist-Worn Accelerometry in the U.K. Biobank. *Sensors (Basel)* **21**(2021).
15. Mirelman, A.*, et al.* Arm swing as a potential new prodromal marker of Parkinson's disease. *Mov Disord* **31**, 1527-1534 (2016).
16. Bycroft, C.*, et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
17. Doherty, A.*, et al.* Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLoS One* **12**, e0169649 (2017).
18. Parkinson’s UK. The incidence and prevalence of Parkinson’s in the UK. Vol. 2022 (2017).

19. Walmsley, R.*, et al.* Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. *Br J Sports Med* **56**, 1008-1017 (2021).
20. Collins, G.S., Reitsma, J.B., Altman, D.G. & Moons, K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* **13**, 1 (2015).
21. Davis, J. & Goadrich, M. The Relationship between Precision-Recall and ROC Curves. in *Proceedings of the 23rd International Conference on Machine Learning* 233–240 (Association for Computing Machinery, 2006).
22. Darweesh, S.K.*, et al.* Trajectories of prediagnostic functioning in Parkinson's disease. *Brain* **140**, 429-441 (2017).
23. Yang, Y.*, et al.* Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nat Med* **28**, 2207-2215 (2022).
24. Swaddiwudhipong, N.*, et al.* Pre-diagnostic cognitive and functional impairment in multiple sporadic neurodegenerative diseases. *Alzheimers Dement* (2022).
25. Chandrasekaran, R., Katthula, V. & Moustakas, E. Patterns of Use and Key Predictors for the Use of Wearable Health Care Devices by US Adults: Insights from a National Survey. *J Med Internet Res* **22**, e22443 (2020).
26. Simon, D.A., Shachar, C. & Cohen, I.G. Unsettled Liability Issues for "Prediagnostic" Wearables and Health-Related Products. *JAMA* **328**, 1391-1392 (2022).
27. Xu, S., Kim, J., Walter, J.R., Ghaffari, R. & Rogers, J.A. Translational gaps and opportunities for medical wearables in digital health. *Sci Transl Med* **14**, eabn6036 (2022).
28. Parkinson Progression Marker, I. The Parkinson Progression Marker Initiative (PPMI). *Prog Neurobiol* **95**, 629-635 (2011).
29. All of Us Research Program, I.*, et al.* The "All of Us" Research Program. *N Engl J Med* **381**, 668-676 (2019).
30. Brigo, F., Matinella, A., Erro, R. & Tinazzi, M. [(1)(2)(3)I]FP-CIT SPECT (DaTSCAN) may be a useful tool to differentiate between Parkinson's disease and vascular or drug-induced parkinsonisms: a meta-analysis. *Eur J Neurol* **21**, 1369-e1390 (2014).

Material and Methods

An overview of the performed analyses and included participants can be found in Figure 1.

Study Population

The UKBB holds in-depth information on ~500,000 participants and is approved by the Research Ethics Committee (reference 16/NW/0274). It was accessed under the application code 69610 with data released to Cardiff University. Written informed consent of all participants was obtained by UKBB. We explored PD and related disorders, namely 'AllCauseDementia', 'AllCauseParkinsonism', 'AlzheimerDisease', 'Dystonia', 'Osteoarthritis', and 'Depression'. We identified patient groups based on ICD10 and ICD9 codes in the hospital inpatient data (fields 41270 and 41271) and the death registry (fields 40001 and 40002) which were curated from UKBB provided tables and phecodes [1], as well as self-reported diagnoses (field 20002). Primary care data was also included (field 42040) using read codes (version 2 and 3) respective to the ICD10 codes as mapped through TRUD NHS Read browser [2]. The respective codes for each diagnosis can be found in Supplementary Table 2. We distinguished prodromal (incident) and diagnosed (prevalent) cases at the date of accelerometer data collection (field 90003) based on the earliest reported date across all resources and allowed for a two-year margin of error, meaning that patients diagnosed before or within the two years after accelerometer data collection are classified as diagnosed/prevalent cases (Figure 3A) and individuals receiving a diagnosis >2 years later are labelled prodromal/incident cases. Unaffected controls were defined as having no neurological, behavioural disorder or any of the included disorders across all included sources and not having been prescribed Antiparkinsonism drugs by a GP (field 42039) (Supplementary Table 17) or self-reporting (field 20003) usage of Antiparkinsonism drugs (ATC = N04, mapped using Supplementary Table 1 of Wu, et al. [3]). From the set of unaffected controls, we randomly sampled unique age- and sex-matched individuals to our PD patients (1:1). Only participants who passed quality control for the accelerometer data (field 90016) were included. Health-related outcome data is

available up to March 2021. Using the same approach, we also identified participants with recognised prodromal signs and symptoms, namely depression, anxiety, orthostatic hypotension, RBD, hyposmia, urinary incontinence, and constipation. We defined these as prodromal symptoms if they were reported before a PD diagnosis was made.

We differentiated between three control groups: (i) matched unaffected controls with 1:1 sex- and age-matching, (ii) all unaffected individuals, and (iii) a representative sample of the general population including all unaffected controls and individuals diagnosed with other disorders such as dementia, dystonia, osteoarthritis, and other forms of parkinsonism. Further details of each of these groups are provided in the 'Prediction Models' section of the methods.

Accelerometer data

103,712 participants who agreed to participate after random email recruiting wore an Axivity AX3 wrist-worn triaxial accelerometer on their dominant hand for a 7-day-period. UKBB provides summary statistics (category 1009) describing daily and hourly averages. We augmented the accelerometer data by pre-processing the raw data into time-series data and classifying 30 second intervals into physical activity categories, namely imputed, sleep, sedentary, light, or Moderate to Vigorous Physical Activity (MVPA), with a machine-learning model, using balanced random forests with Markov confusion matrices, using the accelerometer package provided by the Oxford Wearables Group [4]. The pre-processing steps we employed are the same as those used to derive the summary statistics from UKBB. These steps include device calibration, resampling to 100 Hz, and removal of noise and gravity [5]. From this machine learning labelled time-series data we derived measures of uninterrupted duration, mean movement, and number of interruptions for each physical activity category. For sleep this entailed measures of sleep quality, for example frequency of night-time waking and frequency of daytime napping. For this, we only used complete datasets, so for each participant incomplete hours or days, respective to the measure of interest, were removed. We retained for each measure the highest possible amount of data leading to different data being

used for different summary statistics. For example, maximum hours spent continuously in one physical activity class was calculated over all data available whereas the mean hours per 24 hours spent in one activity class was calculated as the mean of only fully covered 24h periods starting at 10am. 10am was chosen as the offset as data collection was implemented to start at that time. This removal of data was performed to avoid biases through higher representation of specific hours of the day. A list of all calculated derived physical activity and sleep phenotypes can be found in Supplementary Table 16.

Additional data

We merged the accelerometry summary statistics (category 1009), blood biochemistry measures at initial visit (category 17518), physical health measures at initial visit (category 100006), Polygenic Risk Scores (PRS) (category 301), and our derived physical activity phenotypes. We further included age at accelerometer data collection and sex.

Statistical Analyses

Data retrieval from UKBB was facilitated with an adapted version of the ukbb_parser [6] (https://github.com/aschalkamp/ukbb_parser). Data processing and model training were carried out in python 3.8 using scipy 1.6.1, pingouin 0.5.1 [7], scikit-learn 0.23.2 [8] and sksurv 0.14.0 [9] packages. Statistical analyses were performed and figures were generated with python 3.9 pingouin (0.5.1) (Vallat, 2018), seaborn 0.12.1 (Waskom, 2021), and matplotlib 3.6.2. All python environments are provided in the associated github repository.

Prevalence

We validated our established cohort of PD cases by comparing the observed to the expected prevalence. For each year between 1950 and 2021 we identified the number of diagnosed and undiagnosed cases in each age-group. Based on the date of death (field 40007) participants

were excluded from the statistics from their year of death onwards. We calculated the estimated number of PD cases for each year based on the number of people alive in each age group and the prevalence rates for individual age groups from a population-based study from 2015 [10]. We extrapolated this expected prevalence until 2030, making the assumption of no deaths taking place.

Identifying and Adjusting for Covariates

Age and sex are known covariates of acceleration. To address this, we subsampled the unaffected controls in an age- and sex-matched manner. However, prodromal and clinically diagnosed groups differed significantly in age (t-statistic = 3.18, p-value = 1.6×10^{-3}). BMI is also a covariate for acceleration (Supplementary Table 4). To address this, we calculated the residuals of average acceleration (field code = 90087) using coefficients for age, BMI and sex learned from the unaffected control group (N = 36082) with a linear regression model including an intercept. This resulted in the removal of some participants due to missing information of some covariates. When examining the other (non-PD) diagnostic groups, we removed any cases in which comorbid PD was observed to attempt to maintain cohort homogeneity (Supplementary Figure 1). We also removed participants with a comorbid diagnosis of depression from all other diagnosis groups as this diagnostic group was found to have a significantly reduced acceleration and is a prodromal marker for PD. We compared the residual average acceleration measure between the prodromal and diagnosed groups for each included diagnosis class with two-sided T-tests and Bonferroni-correction and for comparisons including healthy controls with the Welch t-test and Bonferroni-correction as here the sample sizes differ substantially between groups. We also computed the residual sleep features, which were age-, BMI- and sex-corrected, using the same method as described above for average acceleration.

Medication effect on acceleration

W used the primary care prescription records (field code = 42039) to identify participants that i) were ever prescribed medication typically used in the treatment of PD and parkinsonism AND ii) had received a prescription for it within 10 weeks before data collection, and hence were likely to be medicated during data collection. Read-codes were taken from the UKBB documentation (https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=594: Parkinson's disease – P2 (all codes referring to medication usage)). A total of 513 UKBB participants had ever been prescribed PD medication. 302 of these were not in the 'AllCauseParkinsonism' group, 206 were among the identified PD cases. 5.37% of the 3837 PD cases in UKBB have been prescribed PD medication at least once. GP information was missing for 2045/53% identified PD cases.

We matched the GP records to the accelerometry data using the date of issue and date of accelerometry collection information. Of the PD cases who have accelerometry data available 20 were ever prescribed medication. 19 of those are diagnosed PD and 1 prodromal PD at accelerometry data collection. 6 of these, including the prodromal case, were prescribed medication only after accelerometry data collection (mean days from accelerometry collection to prescription: 464.42±257.28) and were hence assumed to be not medicated during data collection. 14 of the PD cases taking medication were prescribed medication before accelerometry data collection (mean days since prescription to data collection: 104.56±333.1) and of these 13 subjects had not more than 10 weeks between prescription and data collection (mean days from medication prescription to accelerometry data collection: 15.57±9.85). We compared treated (N = 13) and untreated (N = 122) diagnosed PD cases in terms of average acceleration and found no significant differences, potentially due to small sample sizes. We repeated this for residual average acceleration corrected for age, sex, and BMI which lead to the same results (treated N = 10, untreated N = 103) (Supplementary Figure 4).

Prediction Models

An overview of the trained models with outcome definition and included features can be found in Extended Figure 1. To quantify the predictive power of the acceleration data on an individual level and to compare it to other modalities, we fitted logistic regression models. We identified five modalities: genetics, lifestyle, blood biochemistry, prodromal symptoms, and accelerometry (Table 1). For each modality we included the most recent information available, meaning that for the lifestyle and blood modality we included the features from the initial visit as those were not collected later on, and for the prodromal symptoms we checked up until March 2021 (last update of linked clinical records) for the existence of prodromal symptoms preceding a diagnosis of PD. We restricted the dataset to participants with information available for all five modalities. We estimated the predictive performance of each modality with logistic regression with fitted least absolute shrinkage and selection operator (LASSO) penalty in a nested cross-validation. We chose LASSO to increase sparsity in our model and thus decrease complexity such that the models would be more stable and less prone to overfitting. Logistic regression being one of the simplest algorithms for binary classification tasks was chosen due to its high interpretability and prominence. Three different model types were trained: 1) diagnostic biomarker: identifying diagnosed PD (N = 153) from control, 2) prodromal marker: identifying prodromal PD (N = 113) from control, 3) screening: identifying diagnosed and prodromal PD (N = 266) from control. The control group was either 1:1 sex- and age-matched unaffected controls, all unaffected controls (N = 24987) or a representation of the general population (N = 33009), which included unaffected controls and participants diagnosed with other disorders such as dementia, dystonia, osteoarthritis, and other forms of parkinsonism (N = 8022, participants with comorbidities were only included once). We did not include participants with a single diagnosis of Depression into the control group as the presence of depression before diagnosis of PD was included as a predictor.

We trained models on different modalities, always including the covariates age and sex and an intercept: no-skill (only intercept), genetics, lifestyle, blood, prodromal symptoms, all

acceleration features, all modalities combined (Table 1). We further trained models combining the features of each modality with all accelerometry features. We trained the models in a nested 5-fold cross-validation using a stratified 5-Fold split for both the inner and outer split such that in each fold 20% of the data were used for testing and 80% for the inner fold 20% for validation with grid search for the best Lasso penalty hyperparameter (10 equidistant values between 10^{-5} and 10^{4}). The no-skill model using only an intercept, was trained using only outer cross-validation folds as no hyperparameters had to be fitted in the inner fold, as no penalty was applied here. Parameter selection was applied independently to each training fold. Real valued predictors were standardised based on the training data of the outer split to have a standard deviation of one and a mean of zero. Binary data was encoded as 0/1. The sample size (class imbalance), mean, and standard deviation for each included feature for every case and control group are given in Supplementary Table 9. Balanced class weighting was applied to adjust for class imbalances. We report the mean and 95% confidence interval (CI) of the area under the receiver operator curve (AUROC) and the area under the precision and recall curve (AUPRC) on the outer cross-validation splits to compare models (Supplementary Table 5).

Performance of the classifiers was compared using two-sided T-tests with multiple testing accounted for using Bonferroni correction at 0.05 (Supplementary Table 6). We compared each modality to the accelerometry modality (Figure 5A-C). We further compared each modality to the no-skill performance and each modality with its single performance and its performance when combined with the accelerometry modality (Supplementary Figure 5 & 6). For the accelerometry model we compare its single performance to that of a model using all modalities. We showed the mean AUROC and AUPRC curves with 1 standard deviation across folds for each model (Supplementary Figure 7 – 12). We determined which disorders were most likely to be misidentified by the models as PD by assessing the mean predicted probability of having PD as assigned by the model on the outer folds of the test data (Extended Figure 7). We further investigated feature importance by calculating the mean effect of each predictor

over the five outer cross-validation splits. We validated their stability through checking their effect size in each outer cross-validation split. A feature was labelled as important and significant if the mean effect size across folds was significantly different from zero (95% Bonferroni-corrected CI does not cross zero). In addition to the combined model which used the union of the features across all modalities, we further trained a stacked model which took the predicted probabilities from each modality-specific model and integrated it in a final lasso logistic regression model that predicts the overall probability of having or getting a PD diagnosis. Each modality was thus assigned a coefficient with which its prediction contributed to the final prediction. This model was trained on the same outer cross-validation splits as the other models using their predictions on the training data to train the final model and the respective test data for testing. No penalty hyperparameter was fitted here, so no inner cross-validation was performed. The performance was evaluated in a similar fashion with AUROC and AUPRC across folds and the assigned coefficients across folds were evaluated for their stability across folds by examining the mean and standard deviation (Supplementary Table 5).

Finally, we performed calibration analysis, investigating the calibration at large, the calibration slope, and the calibration curves [11]. The models trained with matched controls achieved good calibration with slopes close to 1 (Supplementary Table 18). The models demonstrating high class imbalance, i.e. those trained with all unaffected controls and the general population, showed poor calibration as they overestimated the prevalence (Extended Figure 9, Supplementary Figure 19). This can be explained through the applied class weighting during training whereby the model was penalized to a greater degree when a case was not detected, compared to when a control was predicted to be a case, thus introducing a bias. Of note, we chose class weighting over calibration, and thus sensitivity over specificity, due to our aim of developing a screening tool rather than one that provides a clinical diagnosis.

Survival Models

We explored the value of each modality to predict the time to diagnosis for the prodromal PD cohort. We did so by first calculating the Pearson correlation (scipy.stats.pearsonr with two-sided) of the residual average acceleration (age- and sex-corrected) and time to diagnosis in the prodromal PD cohort. A simple linear association was not found between residual average acceleration (age- and sex- corrected) and time to diagnosis for the prodromal cases ($r = 0.11$, p-value = 0.13), i.e. average acceleration did not appear to decline further closer to the date of diagnosis. We then used survival modelling on the prodromal (N = 113) and controls (matched: N = 113, unaffected controls: N = 24987, or population: N = 33009) to predict when each individual would be diagnosed (Extended Figure 8). To this end, we used survival random forests with a five-fold stratified cross-validation. The survival random forest is made up of 1000 trees which requires at least 10 samples for a split and 15 samples per leaf. The controls were modelled as right censored, as we did not know whether or when they would receive a diagnosis. We modelled the time from accelerometer data collection to PD diagnosis, thus defining the time of accelerometer collection as time 0. For the prodromal symptoms modality, we hence restricted the time of diagnosis of prodromal symptoms to the date of accelerometer data collection and removed all subsequent diagnoses of prodromal symptoms. The sample size (class imbalance), mean, and standard deviation for each included feature for every case and control group are given in Supplementary Table 9. We reported the time-dependent AUROC (Figure 5D-F) and brier score (Supplementary Figure 20) on the five cross-validation test sets. These metrics are calculated by defining the cases and controls dynamically for several time points with controls transitioning to cases at the time of their PD diagnosis. At each time point based on the current case and control assignment, the predicted case/control assignment is assessed with a standard AUROC using the true positive (sensitivity) and false positive rate (1 – specificity).

Data Availability

Data from the UK Biobank (ukbiobank.ac.uk/) are available to researchers on application to the UK Biobank following the steps outined here: https://www.ukbiobank.ac.uk/enable-your-research.

Code Availability

Code that supports the findings of this study is available on GitHub https://github.com/aschalkamp/UKBBprodromalPD.

References

1. Wu, P.*, et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325 (2019).
2. NHS, D.T. NHS Read Browser. Vol. 2022.
3. Wu, Y.*, et al.* Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat Commun* **10**, 1891 (2019).
4. Willetts, M., Hollowell, S., Aslett, L., Holmes, C. & Doherty, A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Sci Rep* **8**, 7961 (2018).
5. Doherty, A.*, et al.* Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLoS One* **12**, e0169649 (2017).
6. Brandes, N., Linial, N. & Linial, M. PWAS: proteome-wide association study-linking genes and phenotypes by functional variation in proteins. *Genome Biol* **21**, 173 (2020).
7. Vallat, R. Pingouin: statistics in Python. *Journal of Open Source Software* **3**, 1026 (2018).
8. Pedregosa, F.*, et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825—2830 (2011).
9. Pölsterl, S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research* **21**, 1–6 (2020).
10. Parkinson's UK. The incidence and prevalence of Parkinson's in the UK. Vol. 2022 (2017).
11. Van Calster, B.*, et al.* Calibration: the Achilles heel of predictive analytics. *BMC Med* **17**, 230 (2019).