Rapid dynamic naturalistic monitoring of bradykinesia in Parkinson's disease using a wrist-worn accelerometer

Jeroen G.V. Habets*[1], Christian Herff[1], Pieter L. Kubben[1], Mark L. Kuijf[2], Yasin Temel[1], Luc J.W. Evers[3], Bastiaan R. Bloem[3], Philip A. Starr[4], Ro'ee Gilron[4]** & Simon Little[4]**
** Authors contributed equally

1 Maastricht University, Department of Neurosurgery, School of Mental Health and Neuroscience, Maastricht, Netherlands,
2 Maastricht University, Department of Neurology, School of Mental Health and Neuroscience, Maastricht, Netherlands,
3 Radboud university medical center; Donders Institute for Brain, Cognition and Behaviour; Department of Neurology; Center of Expertise for Parkinson & Movement Disorders, Nijmegen, the Netherlands,
4 University of California San Francisco, Department of Movement Disorders and Neuromodulation, San Francisco, United States

***Corresponding author:** Jeroen Habets, Department of Neurosurgery, Maastricht University Medical Center, 6229 ER, Maastricht, The Netherlands. +31433876052, j.habets@maastrichtuniversity.nl.

Key words
Parkinson's disease, naturalistic motor monitoring, wearable sensors

Word count
Main text 2916. Total incl. Methods 4843.

Financial disclosures, conflict of interest
None of the authors have a conflict of interest. JH received funding from the Dutch health care research organization ZonMW (Translational Research 2017 - 2024 grant nr. 446001063). JH, PT and YT received funding from the Stichting Weijerhorst. CH received a VENI-funding from the Dutch Research Council (WMO). SL and research reported in this publication was supported by the National Institute Of Neurological Disorders And Stroke of the National Institutes of Health under Award Number K23NS120037. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Abstract

Introduction: Motor fluctuations in Parkinson's disease are characterized by unpredictability in the timing and duration of dopaminergic therapeutic benefit on symptoms including bradykinesia and rigidity. These fluctuations significantly impair the quality of life of many Parkinson's patients. However, current clinical evaluation tools are not designed for the continuous, naturalistic (real-world) symptom monitoring needed to optimize clinical therapy to treat fluctuations. Although commercially available wearable motor monitoring, used over multiple days, can augment neurological decision making, the feasibility of rapid and dynamic detection of motor fluctuations is unclear. So far, applied wearable monitoring algorithms are trained on group data. Here, we investigate the influence of individual model training on short timescale classification of naturalistic bradykinesia fluctuations in Parkinson's patients using a single wrist-accelerometer.

Methods: As part of the Parkinson@Home study protocol, 20 Parkinson patients were recorded with bilateral wrist-accelerometers for a one hour OFF medication session and a one hour ON medication session during unconstrained activities in their own homes. Kinematic metrics were extracted from the accelerometer data from the bodyside with the largest unilateral bradykinesia fluctuations across medication states. The kinematic accelerometer features were compared over the whole one-hour recordings, and medication-state classification analyses were performed on one-minute segments of data. The influence of individual versus group model training, data window length, and total amount of training patients included in group model training on classification was analyzed.

Results: Statistically significant areas under the curves (AUCs) for medication induced bradykinesia fluctuation classification were seen in 85% of the Parkinson patients at the single minute timescale using the group models. Individually trained models performed at the same level as the group trained models (mean AUC both 0.70, +/- respectively 0.18 and 0.10) despite the small individual training dataset. AUCs of the group models improved as the length of the feature windows was increased to 300 seconds, and with additional training patient datasets.

Conclusion: Medication induced fluctuations in bradykinesia can be classified using wrist worn accelerometery at the time scale of a single minute. Rapid, naturalistic Parkinson motor monitoring has important clinical potential to evaluate dynamic symptomatic and therapeutic fluctuations and help tailor treatments on a fast timescale.

Introduction

Parkinson's disease (PD) is a disabling neurodegenerative disorder characterized by motor and non-motor symptoms that affect patients' motor performance and quality of life (QoL) [1–3]. Symptomatic PD management initially focuses on dopamine replacement therapies [4]. However, half of PD patients develop 'wearing-off' motor fluctuations during the first decade after diagnosis [5,6]. Wearing-off motor fluctuations are defined as inconsistent therapeutic benefits on symptoms such as bradykinesia and rigidity, despite regular dopaminergic delivery [6]. These motor fluctuations and other dopaminergic related side effects can markedly impair patients' QoL [7]. Motor fluctuations are therefore a primary indication for consideration of deep brain stimulation (DBS) [1,8]. Adequate monitoring of motor fluctuations is essential for treatment evaluation, both in the presence and absence of DBS, and wearable motion sensing represents an appealing approach to support this quantification [9,10], although several challenges remain to be addressed [11,12].

Ideally, objective motor fluctuation monitoring should accurately measure and decode movement, during real world (naturalistic) activities, and be simple to implement for patients [10,13]. Currently used Parkinson's evaluation tools such as the Movement Disorders Society Unified Parkinson Disease Rating Scale (MDS-UPDRS) and the Parkinson Disease QoL questionnaire (PDQ-39) are not designed for chronic dynamic, naturalistic symptom monitoring [14,15]. They contain questionnaires which capture subjective estimates of retrospective symptoms over a week (MDS-UPDRS II and IV), or a month (PDQ-39), and these are dependent on patient recall, which is often imperfect, particularly in patients with cognitive dysfunction. Observing and scoring motor fluctuations requires trained health providers to perform single time point evaluations (MDS-UPDRS III). Motor diaries, often used as gold standard for 24-hour naturalistic monitoring, require self-reporting every 30 minutes [16]. This burden causes recall-bias and diary fatigue in long-term use [17].

The strong clinical need for continuous symptom tracking, together with the wide availability and presence of affordable accelerometer-based sensors, has led to several academic and commercially available wearable sensor PD monitoring systems [18–22]. Motor fluctuation monitoring with commercially available devices is currently mostly based on summary metrics derived from multiple days of sensor data and incorporation of these metrics during neurological consultation has led to promising augmentation of clinical decision making [21,23–26]. However, these sensor monitoring systems have thus far been found to be better correlated with clinical PD metrics on a time scale of days rather than hours [21,27], which is a notably longer time window than used in the original development studies [19,21,28–30]. Successful motor fluctuation classification over short time periods (minutes to hours) would enable dynamic therapeutic motor response monitoring, and we suggest individual model training as a methodological improvement to pursue this. So far, motor monitoring algorithms have typically been trained on group data, and individual model training is suggested due to intersubject heterogeneity of PD symptomatology [11,18,20,31]. This hypothesis is strengthened by a recent successful algorithm-innovation combining short and long time epochs in a deep learning model correlating wrist- and ankle-accelerometer metrics with total UPDRS III scores on 5-minute epochs [32].

Here, we investigate the performance of machine learning classification models identifying rapid (single minute level), medication-induced motor fluctuations in PD patients. The classification models are trained on unconstrained naturalistic (at home) motion data derived

from a unilateral wrist-worn accelerometer. Classification models based on individual data are compared with models based on group data. Further, we analyse the influence of the number of individuals included in the group model training data, and the length of analysed accelerometer data epochs (time window lengths), on classification results. We focused symptom decoding on bradykinesia since this cardinal feature of PD and has been found to be more challenging to detect with motion sensors than tremor or dyskinesia [1,33]. This is likely due to higher distributional kinematic overlap of bradykinesia fluctuations with normal movements and normal periods of rest [18,34–36].

We hypothesized that single minute bradykinesia classification would be achievable using machine learning and that individualized motion classification models in PD would demonstrate more reliable short-term classification of naturalistic bradykinesia fluctuations compared to group models.

Results

Study population and recorded data

20 PD patients from the Parkinson@Home data repository [37] were included in this study. The Parkinson@Home study recorded accelerometer data in two medication-states, while PD patients were encouraged to perform an hour of their unconstrained activities in their own homes. MDS-UPDRS III and AIMS scores were assessed immediately prior to their recordings, performed by trained physicians directly in the patients' homes. The first MDS-UPDRS and AIMS assessment took place in the dopaminergic deprived state (pre-medication), and the second assessment was repeated post-medication after patients reported experiencing a full effect of their usual dopaminergic medication. We included PD patients who showed an improvement in the sum of unilateral MDS-UPDRS III items representing upper extremity bradykinesia, in at least one body side (see Methods section). Wrist-accelerometer data only from the side with the largest unilateral upper extremity bradykinesia improvement were included for feature extraction and classification analyses (see Figure 1).

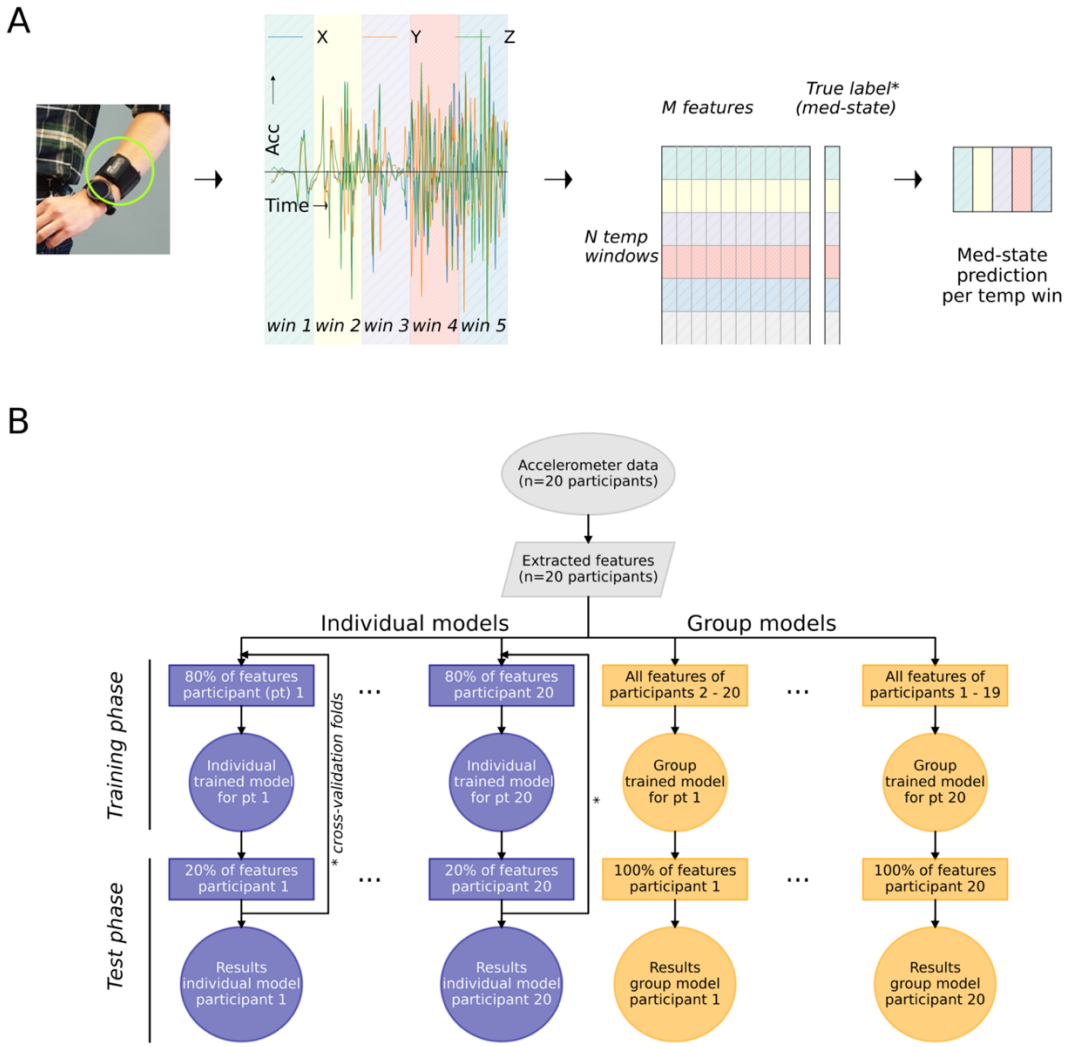

Figure 1: Accelerometer-based Parkinsonian motor fluctuation detection workflows.
A: Left: A wrist-worn motion sensor (Physilog 4, Gait Up SA, CH; green circled) is used to collect unilateral tri-axial accelerometer data. X, Y, and Z represent acceleration (meters/second per second) in the three axes over time (seconds). Temporal windows are determined for data analysis and are indicated in different colours over time (win1, win2, …). Center: Signal preprocessing and feature extraction convert the raw tri-axial signal into a dataset containing M features (Table S1), calculated for every temporal window (in total M columns and N rows). For the training phase of the machine learning classification models, the true labels representing medication states are used. Right: In the testing phase, inserting the feature set (M x N) in the trained classification model leads to N medication state predictions.
B: Workflow to train and test individual and group models. Identical features were extracted from the raw accelerometer data (grey symbols) for every individual participant. For the individually trained models (blue), the features from 80% of a participant's epochs were used in the *training phase* (y-axis). The trained individual model was tested with the remaining, unused, 20% of epochs during the *test phase*. The arrows (*) from test phase to training phase represent the multiple cross-validation folds applied to train and test the individual models on different selections of training and test data. For the group models (yellow), each participant was tested in turn, with data from the other 19 participants used in the *training phase*.

Demographic and disease specific characteristics are presented in Table 1. In total, 3138 minutes of accelerometer data were recorded in the 20 included patients. After balancing the data sets for medication status (see Methods), 2380 minutes of accelerometer data were included. On average 59.5 (+/- 14.3) minutes of accelerometer data from both pre- and post-medication recordings were included per participant. We extracted multiple features which are described in the current literature to index naturalistic bradykinesia with a wrist-accelerometer (see Table S1 for details and references). In total, 103 motion accelerometer features were extracted for every feature window, including both time domain and spectral features from the accelerometer.

Characteristics	
Total number (% female)	20 (60%)
Age (years, mean (sd))	63.4 (6.4)
Accelerometer data per medication state (minutes, mean (sd))	59.5 (14.3)
Accelerometer data per medication state, after activity filtering (minutes, mean (sd))	44.5 (13.9)
PD duration (years, mean (sd))	8.1 (3.5)
Levodopa equivalent daily dosage (milligrams, mean (sd))	959 (314)
MDS-UPDRS III pre-medication MDS-UPDRS III post-medication	43.8 (11.6) 27.1 (9.6)
AIMS pre-medication AIMS post-medication	0.5 (1.8) 3.7 (4.2)

Table 1: Demographic and disease specific characteristics of patient population. AIMS: Abnormal Involuntary Movement Scale. MDS-UPDRS: Movement Disorders Society Unified Parkinson Disease Rating Scale. Sd: standard deviation. *: fluctuation in (sub-)score between pre- and post-medication recording. **: scores related to the upper extremity included in analysis (based on the largest bradykinesia fluctuation).

Group level statistical analysis of cardinal motion features across medication states

First, we compared pre- and post-medication accelerometer recordings at the group level using four accelerometer features which represent commonly used motion features implemented in naturalistic bradykinesia signal processing (maximum acceleration magnitude, coefficient of variation of acceleration magnitude, root mean square of acceleration, and spectral power (below 4 Hz)) [18,38]. The individual mean values per whole medication-state recording were compared at the group level.

The pre- and post-medication recordings significantly differed at the group level based on the individual mean values of the four main accelerometer-features (MANOVA, Wilk's lamba = 0.389, F-value = 14.2, $p < 0.001$). Post-hoc repeated measures ANOVAs demonstrated

that only the individual coefficient of variation averages significantly differed between pre- and post-medication states (p = 0.0042) (figure 2).

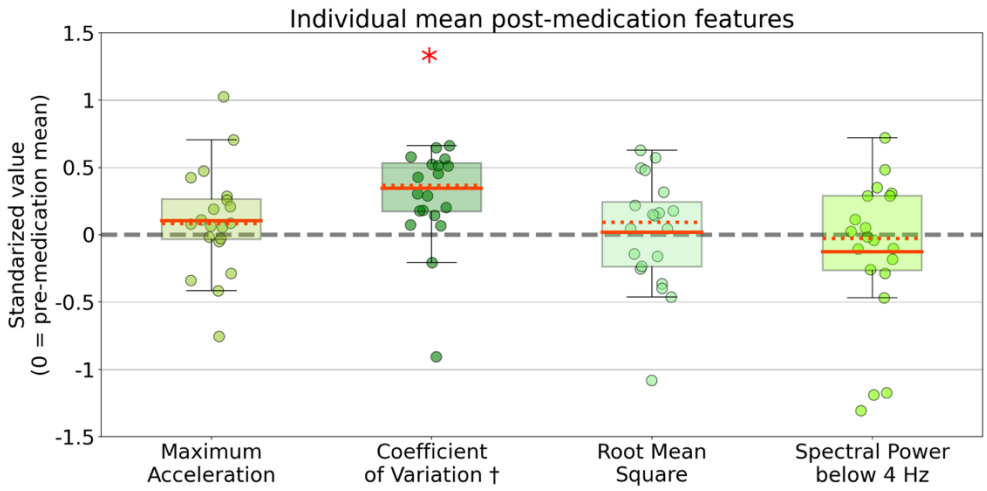

Figure 2: Distributions of individual means for four main movement features. Coloured dots represent the mean feature values during the whole post-medication recording per participant (n=20). Individual post-medication mean values are standardised as z-scores (individual pre-medication recordings are used as references, and therefore the pre-medication mean values equal 0). The red asterix indicates a significant difference on group level between mean coefficient of variations of pre- and post-medication means (alpha=0.05, MANOVA and post-hoc analysis, FDR corrected). † = one positive outlier (1.7) not visualized.

Machine learning classification of short window data epochs

Next, to test classification performance over short time windows (60 second accelerometer feature windows), support vector (SV) and random forest (RF) machine learning models were applied. To tailor the classification models to different activity levels, we repeated all analyses with the inclusion of an “activity filter”, which selected periods of data with at least a minimal amount of movement for analysis (see Methods section). All medication state classifications using the four previously selected motion features led to low AUC scores (means per model ranged between 0.49 and 0.64) and low accuracies (means per model ranged between 49% and 60%) (see table S2 for detailed results).

We therefore repeated our dynamic (1 minute) classification analysis, using an expanded kinematic feature set (103 features). With the full feature set, notably higher AUC scores and classification accuracies were seen for all individual and group (SV and RF) models (table S2, and figures S3AB). Mean AUC scores per model ranged between 0.65 and 0.70, and mean accuracies per model ranged between 60% and 65% (table S2). Most participants yielded AUC scores and accuracies significantly better than chance level (17 out of 20 participants per model), tested through random surrogate dataset generation.

In 90% of participants (18 out of 20) either the best individual or group model classified medication states per 60 seconds significantly better than chance level based on our surrogate datasets (figure 3 and table S2). Group trained models resulted in AUC scores statistically significantly higher than random classification in 17 participants. Individually trained models resulted in AUC scores statistically significantly higher than random classification in 13 participants. Both individual and group models resulted in mean AUC scores of 0.70 (+/- respectively 0.18 and 0.10), and mean accuracies of respectively 65%

(+/- 0.14) and 64% (+/- 0.08) over all 20 participants (figure 3, table S2). Notably, the individual models resulted in a larger standard deviation of AUC scores, including several AUC scores higher then 0.9 as well as below chance level (figure 3).
Overall, these findings confirm the feasibility of rapid naturalistic bradykinesia classification based on wrist-accelerometer metrics. Individual model training resulted in a similar mean AUC with a wider standard deviation compared to group model training.

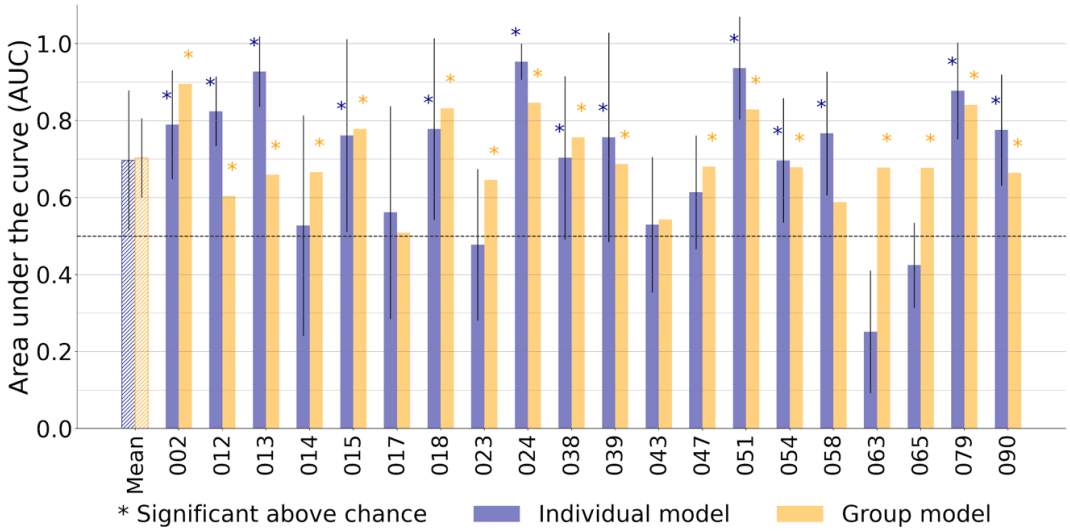

Figure 3: Classification of medication induced motor fluctuations on short accelerometer time windows in individual participants. The first pair of bars represents the mean area under the curve (AUC) score over the twenty participants. Each subsequent pair of bars (002 to 090) represents the AUC scores from one participant. The blue bars represent the AUC score for the individual model, and the yellow bars represent the group model. Note that for the individual models, AUC scores are the averages over the multiple cross-validation folds within a participant (figure 1B). The asterisks indicate whether the corresponding AUC score was significantly better than chance level (5000-repetitions permutation test). Both models have equal mean AUC scores. It is notable that the majority (18 out of 20 of participants) has at least one significant score. Half of the participants yielded a higher AUC score with the individual model than with the group model.

Classification of bradykinesia-centred motor fluctuations versus co-occurring symptoms
We wanted to test whether the models' predictive performance was being driven by confounding fluctuations in tremor or dyskinesia. Therefore, we explored the relation between classification performance and clinical bradykinesia, tremor (both indexed by unilateral upper extremity MDS-UPDRS sub items, see Methods), and abnormal involuntary movement (represented by Abnormal Involuntary Movement Scale (AIMS)) fluctuations. We found no significant correlations between the individual classification performance and the individual clinically scored fluctuations in bradykinesia, tremor, and AIMS (see table S3, all p-values larger than 0.1). At an individual level, we found significant AUC scores in participants with (13, 24, and 79) and without (39, 51, and 58) tremor fluctuations (figure S4A). Similarly, we found significant AUC scores in participants with (2, 15, 51, and 79) and without (39, 18, 24, and 90) AIMS fluctuations (figure S4B).
Individual predictive performance is found not to be proportional to the size of tremor or AIMS fluctuations, which suggests feasibility of using the applied metrics for PD patients with and without tremor and abnormal involuntary movements. Meanwhile, the severity of

bradykinesia did not influence the classification performance, suggesting feasibility of the metrics for patients with even mild-to-moderate bradykinesia fluctuations.

Influence of training data size and feature window length

We next sought to determine the number of patients needed to train a group level model with good classification performance. We found an increase in the predictive performance (AUC) of the group models as the number of patient datasets used during model training was increased (figure 4A). Above 15 participants the increase in mean AUC levelled off towards the 19 included participants.

Next, we also wanted to investigate the impact of the accelerometer data feature window length on the predictive performance of the group models. Increasing the length of the feature windows up to 300 seconds improved the mean AUC (figure 4B). Due to data size limitations, the feature windows were not expanded further than 300 seconds. These analyses could not be reproduced for the individual models due to data size limitations.

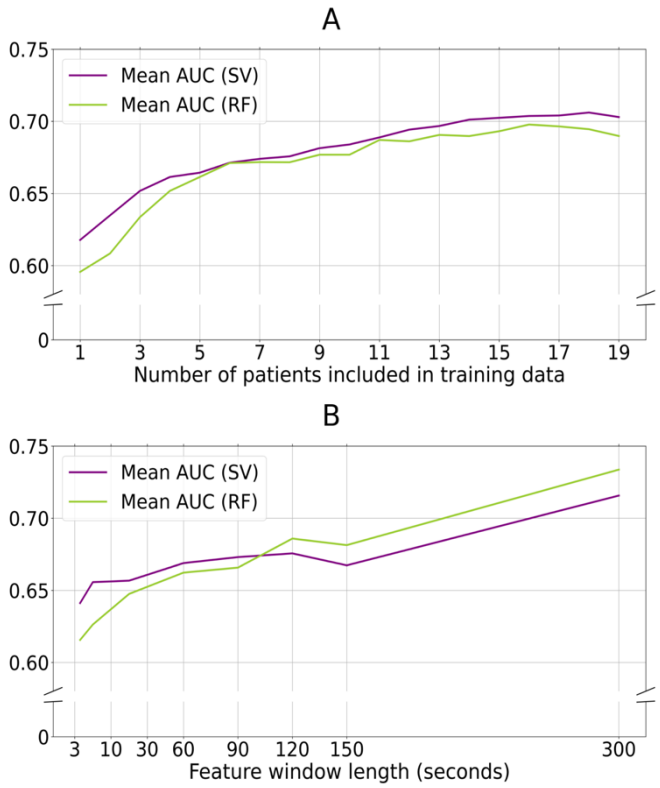

Figure 4: Increasing number of training patients and length of data window duration improves classification performance.
A: Group models are trained for every patient with a varying number of included training data, (x-axis). On the y-axis, the AUC is shown for both SV and RF models (both included activity filtering). An increase in AUC is seen for SV and RF models parallel to an increase in included training patients.
B: Group models are trained for every patient with various feature window lengths (x-axis). On the y-axis, the AUC of the SV and RF models are visualized. Due to the longer feature window lengths, we did not apply the activity filter in this sub-analysis to deal with data size limitations. Larger window lengths up to 300 seconds increased classification performance, while smaller window lengths decrease classification performance.
AUC: area under the receiver operator characteristic; SV: support vector classifier; RF: random forest classifier.

Discussion

Our results demonstrate successful classification of naturalistic bradykinesia fluctuations using wrist-accelerometer data on different timescales using conventional statistical approaches (over one-hour epochs) and machine learning classification (over one-minute epochs). We found that the coefficient of variation of the accelerometer amplitude was significantly increased following dopaminergic medication when a full 60 minutes of data was analyzed per medication condition. At shorter timescales (60 seconds) this feature

(complemented with 3 other accelerometery metrics) was not strongly predictive of medication state using machine learning. However, using a larger number of motion metrics (103), statistically significant classification of medication states could be achieved in 90% of participants (18 out of 20) using either group or the individually trained models (figure 3). Individual and group models resulted both in a mean AUC of 0.70 on the 60 second epochs, where the individual models' AUC scores had a larger standard deviation (figure 3, table S2). Expansion of the data epoch length (from 60 to 300 seconds), as well as inclusion of more training participants, improved AUC scores in the group models. Limited individual data sizes withheld us from testing individual models with expanded data epochs and may explain the larger standard deviation for individual model AUCs (figure 1B and S1).

These results represent the first demonstration of classification of Parkinsonian bradykinesia fluctuations using individually trained models for single wrist-accelerometer data on a rapid timescale. Although we show statistically significant classification over short time windows, we did not find added value yielded by individual model training based on our current results. Reproduction with longer accelerometer recordings for individuals is however likely to improve classification results further.
In general, the presented classification models are notable due to the unconstrained naturalistic (real-world) data collection environment and short time scale of classification. Operating at this shorter timescale, the models show good classification performance compared with benchmark naturalistic medication-state detection models (figure 3) [34–36]. Although better classification performances have previously been described with models using data over longer timescales or from more constrained recordings scenarios, these methodologies improve classification results at the cost of less naturalistic generalizability [18,22,27,28,38]. Adding a second motion sensor can likely increase performance at the cost of user-friendliness and feasibility [32]. Systems using wrist, ankle and/or axial motion sensors have the theoretical advantage of being more sensitive for arm versus leg- / gait centred symptomatology.

Clinical relevance and methodological challenges of naturalistic and rapid PD motor monitoring
Wearable accelerometery based PD monitoring systems have been developed to augment therapeutic decision making, [20,21,23] and to augment clinical assessments in pharmacological trials [11,39]. Previous systems have been validated over the time course of days. However other suggested clinical state tracking applications would require short time scale feedback [22], including fine-grained cycle-by-cycle medication adjustments and conventional [40,41] or adaptive [42–44] [45] deep brain stimulation programming.

Notably, we observed marked differences in classification accuracy using either 4 accelerometer features, or 103 features (figure 3 and S3). This suggests that bradykinesia classification on shorter timescales, requires rich feature sets. The significance testing with surrogate datasets aimed to rule out any resulting overfitting. However, a thorough comparison of feature sets is often complicated by proprietary algorithms or the lack of open-source code [12]. This underlines the importance of transparent, open source, and reproducible movement metric feature sets for naturalistic PD monitoring [46,47].
Another methodological challenge for rapid, objective, naturalistic short-term PD monitoring is the lack of a high quality labelling of data on the same time scale . PD clinical assessment tools, currently applied as gold standards, are limited in their applicability for rapid time

scales. Multiple longitudinal time windows of the dynamic accelerometer time series are labelled with a single clinical score which weakens model training and evaluation. In effect, sensor-based outcomes are often aggregated to match clinical evaluation metrics and time scales which might account for the current upper limit in wearable classification performance [21,24,29]. PD specific eDiaries [48–51], labelled video-recordings on fine time scales [52] and other virtual telemedicine concepts [53] may contribute to this challenge.

Future scientific opportunities to improve naturalistic PD monitoring development

We predict that the coming expansion of real-world motion data sets, containing long-term data over weeks to years in patients with PD, will support optimization of individually trained models [54]. These larger datasets will also allow the exploration of alternative, more data-dependent, computational analyses such as deep neural network classification and learning [35,55]. Moreover, unsupervised machine learning models could also be explored to overcome the issues of lacking temporally matching gold standard for model training and evaluation by surpassing the need of long-term, repetitive, true labels [11,56]. The observed discriminative potential of the coefficient of variation (figure 2) might be of value in post-hoc differentiation of clusters in unsupervised machine learning models.

Additionally, open-source research initiatives should catalyse the development of naturalistic PD monitor models which are not dependent on proprietary software [10,38]. The Mobilize-D consortium for example introduced a roadmap to standardize and structure naturalistic PD monitoring by creating specific 'unified digital mobility outcomes' [46,47]. During the development of these outcomes, features describing distribution ranges and extreme values, rather than means or medians, should be considered [11]. Parallel to open-source initiatives, other creative collaborations between industry and academia such as data-challenges might offer valuable (interdisciplinary) cross-fertilization [57]. Further, adding more limb sensors to improve naturalistic PD monitoring is controversial. Although there is evidence supporting the combined use of wrist, ankle[32,58], or insoles[59] sensory, other reports do not show an improved performance but describe additional burden to the patient [35,60,61].

Limitations

Our study was limited by the individual data set sizes, which restricted inferences that could be made regarding models trained with individual versus group data. Also, the unconstrained character of the pre- and post-medication recordings led to an imbalance in terms of captured activities during the two medication states. The applied activity-filter addressed this limitation partly but does not rule out imbalance in exact activities. This imbalance compromises pattern recognition based data analysis [35], but is also inherent to naturalistic PD monitoring [18] and exploring the boundaries of this limitation is essential for future PD monitor applications. Replication of our methodologies in larger data sets, and inclusion of validated activity classifiers may contribute to overcoming this limitation. Future studies should also aim to detect symptom states beyond a binary differentiation between on- versus off-medication.

Conclusion

We here demonstrate that classification of naturalistic bradykinesia fluctuations at the minute time scale is feasible with machine learning models trained on both individual and group data in PD patients using a single wrist-worn accelerometer. At longer timescales of an hour – a single accelerometer feature, the coefficient of variation, is predictive of bradykinesia at

the group level. Extension of short accelerometer time epochs and an increased number of training patients improved classification of group trained models. Rapid, dynamic monitoring has the potential to support personalized and precise therapeutic optimization with medication and stimulation therapies in Parkinson's patients.

Methods

Study sample

For our analysis, we used data from the Parkinson@Home validation study [37]. Detailed descriptions of the study's protocol and feasibility have been described previously [62,63]. In brief, the study recruited 25 patients diagnosed with PD by a movement disorders neurologist who were all undergoing dopaminergic replacement treatment with oral levodopa therapy. The Parkinson@Home study included PD patients who experienced wearing off periods (MDS-UPDRS part IV item 4.3 ≥1) and had at least slight Parkinson-related gait impairments (MDS-UPDRS part II item 2.12 ≥1 and/or item 2.13 ≥1). Participants who were treated with advanced therapies (DBS or infusion therapies) or who suffered significant psychiatric or cognitive impairments which hindered completion of the study protocol were excluded.

For the current subset of PD patients, we excluded three participants who did not show a levodopa-induced improvement in unilateral upper extremity bradykinesia, on both sides (equal or less than zero points). Unilateral upper extremity bradykinesia was defined as the sum of MDS-UPDRS part III items 3c, 4b, 5b, and 6b for the left side, and items 3b, 4a, 5a, and 6a for the right side. Sum scores from medication on-states were compared with sum scores from medication off-states. For each included participant, only data from the side with the largest clinical change in upper extremity bradykinesia sub items were included. Two participants were further excluded because there was less than 40 minutes of accelerometer data available from their pre- or post-medication recording, resulting in a dataset of 20 patients.

The study protocol was approved by the local medical ethics committee (Commissie Mensgebonden Onderzoek, region Arnhem-Nijmegen, file number 2016-1776). All participants received verbal and written information about the study protocol and signed a consent form prior to participation, in line with the Declaration of Helsinki. The de-identified open source dataset will be made available to the scientific community by the Michael J Fox Foundation.

For our current analysis, only unilateral tri-axial accelerometer data from wrist-worn devices were analysed (Gait Up Physilog 4, Gait Up SA, CH). All data collection was performed in the participants' homes. Recordings consisted of two sessions which took place on the same day. First, the pre-medication recording was performed in the morning after overnight withdrawal of dopaminergic medication. Second, the post-medication recording was performed when the participants experienced the full clinical effect after intake of their regular dopaminergic medication. During both recordings, participants performed an hour of unconstrained activities within and around their houses. At the start of both recordings, a formal MDS-UPDRS III was conducted by a trained clinician.

Data pre-processing and feature extraction

Accelerometer data were sampled at 200 Hz and down sampled to a uniform sampling rate of 120 Hertz (Hz) using piecewise cubic interpolation. The effect of gravity was removed from each of the three time series (x-, y-, and z-axes) separately, by applying a 'l1-trend filter' designed to analyse time-series with an underlying piecewise linear trend [64]. Time series were low-pass filtered at 3.5 Hz to attenuate frequencies typically associated with Parkinsonian tremor in accelerometer time series [65]. In addition to the three individual accelerometer time series, we computed a composite time series containing the vector magnitude of the three individual accelerometer axes [x2 + y2 + z2].

Multiple features previously shown to correlate with bradykinesia were extracted from the four time-series (x, y, z, and vector magnitude) (see extensive overview including references in table S1). The features included characteristics from the temporal domain, such as extreme values, variances, jerkiness, number of peaks, and root mean squares, and the spectral domain, such as spectral power in specific frequency ranges, and dominant frequencies. The standard window length of analysis for each extracted feature was set as 60 seconds, meaning one mean value per feature was extracted per time series over every 60 seconds of data. To explore the influence of varying window lengths (3, 10, 30, 90, 120, 150, and 300 seconds), separate feature sets were extracted for each sub analysis. All individual feature sets were balanced for medication-status by discarding the surplus of available data in the longest recording (pre- or post-medication). Features were standardised by calculating individual z-scores per feature. To not average out pre- and post-medication differences, the mean of only the pre-medication recordings was extracted from a value, and the result was divided by the standard deviation of only the pre-medication recordings [66].

Descriptive statistics and analysis of variance

The demographic and disease characteristics of the included participants are described in Table 1. Unilateral scores are provided only for the side on which accelerometer data was analysed. To first test statistical distinguishability of the pre- and post-medication recordings at the group level, before using the entire dataset as an input, four main accelerometer features were chosen a priori. These four features covered the most often used domains of motion metrics applied for naturalistic bradykinesia monitoring (maximum acceleration, coefficient of variation of acceleration over time, root mean square of acceleration over time, and the total spectral power below 4 Hz) [18,38], and were extracted from the vector magnitude time series. Individual averages of each of the four features over the entire dataset (~60 minutes per condition) were analysed for statistically significant differences between the medication states with a multivariate analysis of variance (M-ANOVA). Post-hoc repeated measures ANOVA were performed to explore which feature(s) contributed to the pre- versus post-medication difference. An alpha-level of 0.05 was implemented and multiple comparison correction was performed using the false discovery rate (FDR) method described by Benjamini and Hochberg [67].

Classification of medication states

Individually trained and group trained models

Supervised classification analyses were performed to test whether differentiation between short-term pre- and post-medication was feasible, based on 60-second accelerometer features (figure 1). First, this was tested using the four previously mentioned features extracted from the vector magnitude signal, afterwards the feature set was expanded to

include all described features, as well as the x, y, z time series (table S1). Analyses were performed using a support vector machine (SV) and a random forest (RF) classifier. Classification models trained on individual data and models trained on group data were then compared (figure 1B).

For individually trained models, 80% of a participant's total balanced data was used as training data, and 20% as test data (figure S1). Small blocks (2%) of training data which neighboured the test data were discarded (figure S1) to decrease the temporal dependence between training and test data. To prevent bias caused by the selected block of test data, a 41-fold cross-validation was performed. Each fold (out of 41) includes two continual blocks of 10% of total data, one block from the pre- and one block from the post-medication recording as test data (percentiles 0 to 10 and 50 - 60, percentiles 1 to 11 and 51 to 61, …, and percentiles 40 to 50 and 90 to 100, see visualisation in figure S2).

For group trained models, a leave one out cross-validation was performed. For every participant, a model was trained based on all data (balanced for medication status) from the remaining 19 participants and tested on all data (balanced for medication status) of the specific participant (figure 1B). To assess all models, the area under the receiver operator curve (AUC) and the classification accuracy were calculated as predictive metrics. For the individual models, individual classification outcomes were averaged over the 41 folds.

To test statistical significance of each individual and group model performance, 5000 permutation tests were performed, in which medication state labels were shuffled. The 95th percentile of permutation scores was taken as significance threshold (alpha = 0.05), and FDR multiple comparison corrections were performed [67].

Activity filtering

The reported analyses were repeated after removing data windows without movement activity. To identify data windows that do not contain any motion activity, different methodologies of activity filtering are described in PD monitoring literature [9,38,68]. We applied an activity filter which classified every 60 seconds window with a coefficient of variation of the vector magnitude less than 0.3 as 'no activity' and discarded them from analysis (figure S2). The choice of selected feature was based on previous work[38], and the threshold is chosen pragmatically by group-level observations of video-annotated sections identified as non-active [37]. The activity-filtered data sets were individually balanced for medication-states. For example, if a participant's data set resulted in 50 'active' minutes pre-medication, and only 45 'active' minutes post-medication, the surplus of features from 5 'active' minutes pre-medication were discarded at the end of the data set. On average, 44.5 minutes (+/- 13.9 minutes) of features were included after applying the activity filter and balancing the individual data to include equal individual features per medication state (Table 1).

The influence of training data size, and feature window lengths

To test the impact of the size of the training set in the group models, the training phases were repeated with varying numbers of participants included in the training data (figure 4). As in the original group model analysis, the test data consisted of all data from one participant. The number of training data participants varied between 1 and 19. To prevent selection bias in the selection of the training participants, the analyses were repeated five times per number of included training participants, with different random selections of training participants. Individual model classification were excluded from this analysis by definition.

To analyse the influence of feature window lengths, we repeated the group model analysis with features extracted from data windows of 3, 10, 30, 90, 120, 150, and 300 seconds duration (figure 4). For every analysis, one participant was selected as a test participant, and the other 19 were training participants. This was repeated for all participants and the averages over 20 test participants were reported. This was performed at the group level modelling only, as individual models were limited by total available data size.

Comparing two models' predictive performance

Equality plots were drawn to compare the AUC scores and accuracies between two models, for example a model using 4 features versus 103 features, a model using a SV classifier versus a RF classifier, a model with versus without activity filtering (figure S3). All comparisons were performed separately for the individual and group models. For example, model A led to a higher AUC score than model B in 14 out of 20 participants (14 dots above the equality line). Permutation tests plotted 20 random dots on an equality plot and tested whether the permuted distribution generated 14 or more dots (out of 20) above the equality line. This was repeated 5000 times, and the probability that the distribution '14 out of 20' was the result of chance was determined.

Predictive performance and clinical assessed symptom fluctuations

The influence of clinical bradykinesia, tremor and abnormal involuntary movement fluctuations on predictive performance was tested at a group level by Spearman R correlations between the fluctuation in individual bradykinesia and tremor sub scores and AIMS scores, and the predictive performance (table S3). Individual participants were visualized according to descending tremor and AIMS fluctuation ratings to enable visual comparison of predictive performance with and without co-occurring tremor and abnormal involuntary movement fluctuation (respectively figure S4A and S4B). The tremor scores consisted of the MDS-UPDRS III items representing unilateral upper extremity tremor (items 15b, 16b, and 17b for the left side, and items 15a, 16a, and 17a for the right side).

Software

Raw acceleration time series were down sampled and filtered (for gravity effects) in Matlab. All further pre-processing, feature extraction, and analysis was performed in Jupyter Notebook (Python 3.7). The code used to extract features and analyse data is available at www.github.com/jgvhabets/brady_reallife/ [69].

Data availability

The de-identified open source dataset will be made available to the scientific community by the Michael J Fox Foundation.

Code availability

The code used to extract features and analyse data is available at github.com/jgvhabets/brady_reallife/.

Acknowledgements

The authors want to thank Robert Wilt for his organizational support during this project. This research was supported by a personal travel grant awarded to JH by the Dutch Science Funding Body ZonMW (Translational Research 446001063) and by the Weijerhorst Foundation grant awarded to YT and PK.

Author Contributions

Study conception and design: JH, SL, RG, PS. Data acquisition: BB, LE. Data analysis: JH, RG, SL. Writing manuscript: JH, SL, RG. Critical revision of analysis and manuscript: all.

Competing Interests statement

The authors declare to have no competing interests.

References

1. Bloem, B. R., Okun, M. S. & Klein, C. Parkinson's disease. *Lancet Lond. Engl.* (2021) doi:10.1016/S0140-6736(21)00218-X.
2. Jankovic, J. & Tan, E. K. Parkinson's disease: etiopathogenesis and treatment. *J Neurol Neurosurg Psychiatry* **91**, 795–808 (2020).
3. Kuhlman, G. D., Flanigan, J. L., Sperling, S. A. & Barrett, M. J. Predictors of health-related quality of life in Parkinson's disease. *Parkinsonism Relat. Disord.* **65**, 86–90 (2019).
4. de Bie, R. M. A., Clarke, C. E., Espay, A. J., Fox, S. H. & Lang, A. E. Initiation of pharmacological therapy in Parkinson's disease: when, why, and how. *Lancet Neurol.* **19**, 452–461 (2020).
5. Fasano, A. *et al.* Characterizing advanced Parkinson's disease: OBSERVE-PD observational study results of 2615 patients. *BMC Neurol* **19**, 50 (2019).
6. Kim, H.-J. *et al.* Motor Complications in Parkinson's Disease: 13-Year Follow-up of the CamPaIGN Cohort. *Mov. Disord.* **35**, 185–190 (2020).
7. Hechtner, M. C. *et al.* Quality of life in Parkinson's disease patients with motor fluctuations and dyskinesias in five European countries. *Parkinsonism Relat. Disord.* **20**, 969–974 (2014).
8. Fox, S. H. *et al.* International Parkinson and movement disorder society evidence-based medicine review: Update on treatments for the motor symptoms of Parkinson's disease. *Mov Disord* **33**, 1248–1266 (2018).
9. van Hilten, J. J., Middelkoop, H. A., Kerkhof, G. A. & Roos, R. A. A new approach in the assessment of motor activity in Parkinson's disease. *J Neurol Neurosurg Psychiatry* **54**, 976–9 (1991).
10. Espay, A. J. *et al.* A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. *Mov Disord* **34**, 657–663 (2019).
11. Warmerdam, E. *et al.* Long-term unsupervised mobility assessment in movement disorders. *Lancet Neurol.* **19**, 462–470 (2020).
12. Fasano, A. & Mancini, M. Wearable-based mobility monitoring: the long road ahead. *Lancet Neurol.* **19**, 378–379 (2020).
13. Odin, P. *et al.* Viewpoint and practical recommendations from a movement disorder specialist panel on objective measurement in the clinical management of Parkinson's disease. *NPJ Park. Dis* **4**, 14 (2018).
14. Goetz, C. G. *et al.* Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov Disord* **23**, 2129–70 (2008).
15. Hagell, P. & Nygren, C. The 39 item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence based medicine. *J Neurol Neurosurg Psychiatry* **78**, 1191–8 (2007).
16. Hauser, R. A., McDermott, M. P. & Messing, S. Factors associated with the development of motor fluctuations and dyskinesias in Parkinson disease. *Arch. Neurol.* **63**, 1756–1760 (2006).
17. Papapetropoulos, S. S. Patient diaries as a clinical endpoint in Parkinson's disease clinical trials. *CNS Neurosci Ther* **18**, 380–7 (2012).
18. Thorp, J. E., Adamczyk, P. G., Ploeg, H. L. & Pickett, K. A. Monitoring Motor Symptoms During Activities of Daily Living in Individuals With Parkinson's Disease. *Front Neurol* **9**, 1036 (2018).

19. Horne, M. K., McGregor, S. & Bergquist, F. An objective fluctuation score for Parkinson's disease. *PLoS One* **10**, e0124522 (2015).
20. Sama, A. *et al.* Estimating bradykinesia severity in Parkinson's disease by analysing gait through a waist-worn sensor. *Comput Biol Med* **84**, 114–123 (2017).
21. Powers, R. *et al.* Smartwatch inertial sensors continuously monitor real-world motor fluctuations in Parkinson's disease. *Sci. Transl. Med.* **13**, eabd7865 (2021).
22. Sica, M. *et al.* Continuous home monitoring of Parkinson's disease using inertial sensors: A systematic review. *PLOS ONE* **16**, e0246528 (2021).
23. Pahwa, R. *et al.* Role of the Personal KinetiGraph in the routine clinical assessment of Parkinson's disease: recommendations from an expert panel. *Expert Rev Neurother* **18**, 669–680 (2018).
24. Santiago, A. *et al.* Qualitative Evaluation of the Personal KinetiGraph TM Movement Recording System in a Parkinson's Clinic. *J. Park. Dis.* **9**, 207–219 (2019).
25. Joshi, R. *et al.* PKG Movement Recording System Use Shows Promise in Routine Clinical Care of Patients With Parkinson's Disease. *Front Neurol* **10**, 1027 (2019).
26. Farzanehfar, P., Woodrow, H. & Horne, M. Assessment of Wearing Off in Parkinson's disease using objective measurement. *J Neurol* (2020) doi:10.1007/s00415-020-10222-w.
27. Ossig, C. *et al.* Correlation of Quantitative Motor State Assessment Using a Kinetograph and Patient Diaries in Advanced PD: Data from an Observational Study. *PLoS One* **11**, e0161559 (2016).
28. Rodriguez-Molinero, A. *et al.* Validation of a portable device for mapping motor and gait disturbances in Parkinson's disease. *JMIR Mhealth Uhealth* **3**, e9 (2015).
29. Rodriguez-Molinero, A. Monitoring of Mobility of Parkinson's Patients for Therapeutic Purposes - Clinical Trial (MoMoPa-EC). NCT04176302 (2019).
30. Great Lake Technologies. Kinesia 360 Parkinson's Monitoring Study. (2018).
31. van Halteren, A. D. *et al.* Personalized Care Management for Persons with Parkinson's Disease. *J. Park. Dis.* **10**, S11–S20 (2020).
32. Hssayeni, M. D., Jimenez-Shahed, J., Burack, M. A. & Ghoraani, B. Ensemble deep model for continuous estimation of Unified Parkinson's Disease Rating Scale III. *Biomed. Eng. Online* **20**, 32 (2021).
33. Clarke, C. E. *et al. UK Parkinson's Disease Society Brain Bank Diagnostic Criteria*. *Clinical effectiveness and cost-effectiveness of physiotherapy and occupational therapy versus no therapy in mild to moderate Parkinson's disease: a large pragmatic randomised controlled trial (PD REHAB)* (NIHR Journals Library, 2016).
34. Shawen, N. *et al.* Role of data measurement characteristics in the accurate detection of Parkinson's disease symptoms using wearable sensors. *J. NeuroEngineering Rehabil.* **17**, 52 (2020).
35. Lonini, L. *et al.* Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *Npj Digit. Med.* **1**, 64 (2018).
36. Fisher, J. M. *et al.* Unsupervised home monitoring of Parkinson's disease motor symptoms using body-worn accelerometers. *Parkinsonism Relat. Disord.* **33**, 44–50 (2016).
37. Evers, L. J. *et al.* Real-Life Gait Performance as a Digital Biomarker for Motor Fluctuations: The Parkinson@Home Validation Study. *J Med Internet Res* **22**, e19068 (2020).
38. Mahadevan, N. *et al.* Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *NPJ Digit Med* **3**, 5 (2020).

39. Khodakarami, H. *et al.* Prediction of the Levodopa Challenge Test in Parkinson's Disease Using Data from a Wrist-Worn Sensor. *Sensors* **19**, 5153 (2019).
40. Vibhash, D. S. *et al.* Telemedicine and Deep brain stimulation - Current practices and Recommendations. *Parkinsonism Relat. Disord.* (2021) doi:10.1016/j.parkreldis.2021.07.001.
41. van den Bergh, R., Bloem, B. R., Meinders, M. J. & Evers, L. J. W. The state of telemedicine for persons with Parkinson's disease. *Curr. Opin. Neurol.* **34**, 589–597 (2021).
42. Velisar, A. *et al.* Dual threshold neural closed loop deep brain stimulation in Parkinson disease patients. *Brain Stimulat.* **12**, 868–876 (2019).
43. Castaño-Candamil, S. *et al.* A Pilot Study on Data-Driven Adaptive Deep Brain Stimulation in Chronically Implanted Essential Tremor Patients. *Front. Hum. Neurosci.* **14**, (2020).
44. Gilron, R. *et al.* Long-term wireless streaming of neural recordings for circuit discovery and adaptive stimulation in individuals with Parkinson's disease. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00897-5.
45. Habets, J. G. V. *et al.* An update on adaptive deep brain stimulation in Parkinson's disease. *Mov Disord* **33**, 1834–1843 (2018).
46. Rochester, L. *et al.* A Roadmap to Inform Development, Validation and Approval of Digital Mobility Outcomes: The Mobilise-D Approach. *Digit. Biomark.* **4**, 13–27 (2020).
47. Kluge, F. *et al.* Consensus based framework for digital mobility monitoring. *medRxiv* 2020.12.18.20248404 (2020) doi:10.1101/2020.12.18.20248404.
48. Heijmans, M. *et al.* Monitoring Parkinson's disease symptoms during daily life: a feasibility study. *NPJ Park. Dis* **5**, 21 (2019).
49. Vizcarra, J. A. *et al.* The Parkinson's disease e-diary: Developing a clinical and research tool for the digital age. *Mov Disord* (2019) doi:10.1002/mds.27673.
50. Habets, J. *et al.* Mobile Health Daily Life Monitoring for Parkinson Disease: Development and Validation of Ecological Momentary Assessments. *JMIR Mhealth Uhealth* **8**, e15628 (2020).
51. Habets, J. G. V. *et al.* A Long-Term, Real-Life Parkinson Monitoring Database Combining Unscripted Objective and Subjective Recordings. *Data* **6**, 22 (2021).
52. Bourke, A. K. *et al.* A Physical Activity Reference Data-Set Recorded from Older Adults Using Body-Worn Inertial Sensors and Video Technology-The ADAPT Study Data-Set. *Sensors* **17**, E559 (2017).
53. Alberts, J. L. *et al.* Use of a Smartphone to Gather Parkinson's Disease Neurological Vital Signs during the COVID-19 Pandemic. *Park. Dis.* **2021**, 5534282 (2021).
54. Bloem, B. R. *et al.* The Personalized Parkinson Project: examining disease progression through broad biomarkers in early Parkinson's disease. *BMC Neurol* **19**, 160 (2019).
55. J. Watts, A. Khojandi, R. Vasudevan, & R. Ramdhani. Optimizing Individualized Treatment Planning for Parkinson's Disease Using Deep Reinforcement Learning. in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* 5406–5409 (2020). doi:10.1109/EMBC44109.2020.9175311.
56. Matias, R., Paixão, V., Bouça, R. & Ferreira, J. J. A Perspective on Wearable Sensor Measurements and Data Science for Parkinson's Disease. *Front. Neurol.* **8**, 677 (2017).
57. MJFF, S. BEAT-PD DREAM Challenge (by Sage Bionetworks; Michael J. Fox Foundation). (2020).
58. Pulliam, C. L. *et al.* Continuous Assessment of Levodopa Response in Parkinson's Disease Using Wearable Motion Sensors. *IEEE Trans Biomed Eng* **65**, 159–164 (2018).

59. Hua, R. & Wang, Y. Monitoring Insole (MONI): A Low Power Solution Toward Daily Gait Monitoring and Analysis. *IEEE Sens. J.* **19**, 6410–6420 (2019).
60. Daneault, J. *et al.* Estimating Bradykinesia in Parkinson's Disease with a Minimum Number of Wearable Sensors. in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* 264–265 (2017). doi:10.1109/CHASE.2017.94.
61. Daneault, J.-F. *et al.* Accelerometer data collected with a minimum set of wearable sensors from subjects with Parkinson's disease. *Sci. Data* **8**, 48 (2021).
62. Lima, A. L. S. de *et al.* Large-Scale Wearable Sensor Deployment in Parkinson's Patients: The Parkinson@Home Study Protocol. *JMIR Res. Protoc.* **5**, e5990 (2016).
63. Lima, A. L. S. de *et al.* Feasibility of large-scale deployment of multiple wearable sensors in Parkinson's disease. *PLOS ONE* **12**, e0189161 (2017).
64. Kim, S. J., Koh, K. ,. Boyd, S. ,. Gorinevsky, D. l(1) Trend Filtering. *SIAM Rev* **May 51**, 339–360 (2009).
65. van Brummelen, E. M. J. *et al.* Quantification of tremor using consumer product accelerometry is feasible in patients with essential tremor and Parkinson's disease: a comparative study. *J Clin Mov Disord* **7**, 4 (2020).
66. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
67. Korthauer, K. *et al.* A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* **20**, 118 (2019).
68. Keijsers, N. L., Horstink, M. W. & Gielen, S. C. Ambulatory motor assessment in Parkinson's disease. *Mov Disord* **21**, 34–44 (2006).
69. jgvhabets/brady_reallife: First release for short-term, individual and group modelling analyses | Zenodo. https://zenodo.org/record/4734199#.YJAOZRQza3J.