Development of machine learning: derived mhealth composite biomarkers for trial@home clinical trials

Zhuparris, A.

PART IV

DETECTION OF TREATMENT EFFECTS

CHAPTER 8

Treatment detection and movement disorder society-unified Parkinson's disease rating scale, part III estimation using finger tapping tasks

Ahnjili ZhuParris, MSc,[1,2,3] Eva Thijssen, MSc,[,1,2] Willem O. Elzinga, MSc,[1]
Soma Makai-Bölöni, MSc,[1,2] Wessel Kraaij, PhD,[3]
Geert J. Groeneveld, MD, PhD[1,2] and Robert J. Doll, PhD[1]

Movement disorders. 2023. doi:10.1002/mds.29520

[1] Centre for Human Drug Research (CHDR), Leiden, NL

[2] Leiden University Medical Centre (LUMC), Leiden, NL

[3] Leiden Institute of Advanced Computer Science (LIACS), Leiden, NL

Abstract

The validation of objective and easy-to-implement biomarkers that can monitor the effects of fast-acting drugs among Parkinson's disease (PD) patients would benefit antiparkinsonian drug development. We developed composite biomarkers to detect levodopa/carbidopa effects and to estimate PD symptom severity. For this development, we trained machine learning algorithms to select the optimal combination of finger tapping task features to predict treatment effects and disease severity. Data were collected during a placebo-controlled, crossover study with 20 PD patients. The alternate index and middle finger tapping (IMFT), alternative index finger tapping (IFT), and thumb–index finger tapping (TIFT) tasks and the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) III were performed during treatment. We trained classification algorithms to select features consisting of the MDS-UPDRS III item scores; the individual IMFT, IFT, and TIFT; and all three tapping tasks collectively to classify treatment effects. Furthermore, we trained regression algorithms to estimate the MDS-UPDRS III total score using the tapping task features individually and collectively. The IFT composite biomarker had the best classification performance (83.50% accuracy, 93.95% precision) and outperformed the MDS-UPDRS III composite biomarker (75.75% accuracy, 73.93% precision). It also achieved the best performance when the MDS-UPDRS III total score was estimated (mean absolute error: 7.87, Pearson's correlation: 0.69). We demonstrated that the IFT composite biomarker outperformed the combined tapping tasks and the MDS-UPDRS III composite biomarkers in detecting treatment effects. This provides evidence for adopting the IFT composite biomarker for detecting antiparkinsonian treatment effect in clinical trials.

Introduction

Parkinson's disease (PD) motor impairments can be characterized as slow and rigid and can lead to a gradual reduction in movement speed over time.[1] The recommended instrument for assessing the severity of PD motor symptoms is the Movement Disorder Society's revised version of the Unified Parkinson's Disease Rating Scale, Part III (MDS-UPDRS III).[2] The MDS-UPDRS III offers a reliable and valid metric for evaluating motor manifestations in each body area affected by PD.[3-5] There are two main limitations of the MDS-UPDRS III. First, the MDS-UPDRS III requires approximately 15 minutes to complete with a trained rater, therefore making it time consuming and labor intensive.[6] Thus, MDS-UPDRS III is not ideal for demonstrating the time of onset of fast-acting dopaminergic drugs, such as the inhaled and intranasal forms of levodopa (L-dopa)/carbidopa and apomorphine.[7,8] Second, the MDS-UPDRS III provides only a coarse rating of motor function and therefore cannot identify or differentiate between specific kinematics of finger movements.[3] As fine motor control abnormalities are typically the first manifestations of motor impairments in PD patients, it is important to develop composite biomarkers that are sensitive to these changes.[9] To address these limitations, there is a demand for biomarkers that detect fine-grained changes in motor function and are congruent with the MDS-UPDRS.

Finger tapping tasks provide insights into fine motor activity [10,11] and have been shown to be quick, effective, and simple assessments for estimating MDS-UPDRS motor disability[12,13] and assessing antiparkinsonian drug effects.[14-19] These tasks provide insights into fin- ger and forearm movement speed, accuracy, amplitude, frequency, rhythm, and fatigue.[10,14,20,21] PD patients often experience tremors, stiffness, and difficulty with movement, which can significantly impact their ability to perform daily activities, including buttoning a shirt, typing on a keyboard, or using utensils.[22,23] As patients want treatments that will improve their ability to carry out daily activities, measuring motor function through tapping biomarkers can provide a more direct and meaningful assessment of the impact of treatments on patients' lives. Therefore, the tapping tasks could be considered of interest to both clinicians and patients.

The complexity of parkinsonism motor impairment manifestations cannot be captured by a single biomarker. By exploiting machine learning algorithms, we can combine multiple objective biomarkers into a single composite biomarker that would represent a multi-dimensional characterization of PD.[24] Previous studies have demonstrated that composite biomarkers could effectively differentiate between PD and healthy controls and estimate MDS-UPDRS III symptom severity.[25-27] This study investigates the accuracy and sensitivity of composite tapping biomarkers to detect drug effects and to estimate disease severity among PD patients.

Patients and Methods

This is an extension of a previous study that investigated the reliability of tapping tasks to detect the longitudinal effects of L-dopa/carbidopa and to determine the correlation of the tapping features with the MDS-UPDRS III.[14] The study was conducted at the Centre for Human Drug Research (CHDR) in Leiden, the Netherlands, between July and November 2020 and is registered in the Netherlands Trial Register (trial NL8617).

STUDY OVERVIEW

We conducted a double-blind, placebo-controlled, randomized, two-way crossover study with L-dopa/carbidopa in 20 PD patients that had recognizable off episodes (symptoms not adequately controlled by their medication).[28] Patients received a semi-individual dose of the investigational drug. To ensure an off-on transition, the patients were given a supramaximal dose that was at least 25% higher than their usually administered morning dose.[29]

PATIENT CRITERIA

Enrolled patients had a clinical diagnosis of PD, as confirmed by a neurologist, and a classification of a Hoehn–Yahr stages I to III during their on state by an investigator. Patients were included if they were between ages 20 and 85 years during screening, experienced self-described motor fluctuations, and were taking oral antiparkinsonian medication. Patients were excluded if they had known conditions that would affect L-dopa/carbidopa treatment or study compliance, such as previous intolerance, drug dependence, or psychiatric disease.

ASSESSMENTS

MDS-UPDRS III We selected the MDS-UPDRS III as the gold standard for the purposes of this study. The MDS-UPDRS III was conducted by trained raters at CHDR. The examination took on average 15 minutes to complete. It was performed pre-dose and at 10, 30, 60, and 90 minutes after dosing.

FINGER TAPPING TASKS All the tapping tasks were performed twice pre-dose and once at 10, 25, 45, 60, 75, 90, and 105 minutes after dosing. If the tapping tasks and MDS-UPDRS III were planned simultaneously, then tapping tasks were performed first.

ALTERNATE INDEX AND MIDDLE FINGER TAPPING AND ALTERNATE INDEX FINGER TAPPING Each patient was provided with a touchscreen laptop equipped with the alternate index and middle finger tapping (IMFT) and alternate index finger tapping (IFT) tasks.[10] The patients were instructed to use the hand that was most affected (if both hands were equally affected, to use their dominant hand) and to perform each task as fast and accurately as possible for 30 seconds. For the IMFT, patients were asked to tap between the two targets (2.5 cm apart) with their index and middle fingers. For the IFT, patients were asked to tap the targets (20 cm apart) with their index finger. The IMFT and IFT require two different movements; the IMFT and IFT are dependent on fine finger and forearm movements, respectively.[10] Each of the two tasks generated 43 features relating to speed (eg, total number of taps), accuracy (eg, spatial error), rhythm (eg, intertap interval), and fatigue (eg, change in velocity) (Table S1).[10,14]

THUMB–INDEX FINGER TAPPING A wireless goniometer (Biometrics Ltd, Newport, UK) was placed on the metacarpal and proximal phalanx of the index finger of the most affected hand (if both hands were equally affected, to use their dominant hand).[10,14,30] Each patient was instructed to sit comfortably, hold up the hand, and tap the index finger on the thumb as widely and quickly as possible continuously for 15 seconds. The thumb–index finger tapping (TIFT) assesses unilateral sequential fine finger movements. The 25 features of the TIFT include progressive changes in amplitude, hesitations, and tapping speed during the task (Table S1).[14]

STATISTICAL ANALYSIS

All data preprocessing and statistical analyses were conducted using Python (version 3.8.0) (31) and the Scikit-Learn library (version 1.0.1).[32]

DATA PREPROCESSING All features were visually and statistically inspected for normality using histograms and Shapiro–Wilk tests, respectively. Log or square root transformations were applied when the features were not normally distributed. Only features that were normally distributed were included in the analysis. Missing values were not imputed, and only complete cases were considered.

As the tapping composite biomarker is designed to be a proxy for overall motor function, we did not account for laterality of the tapping task in the biomarkers. The need for assessing the tapping tasks with both hands is therefore avoided, which could streamline the assessment process and reduce the burden on patients.

COMPOSITE BIOMARKERS We developed 10 composite biomarkers. The composite biomarkers represented the baseline-uncorrected or baseline-corrected MDS-UPDRS III 18-item scores; all three tapping tasks combined; and the IFT, IMFT, and TIFT tasks individually. From a statistical viewpoint, we corrected for baseline to remove any concomitant variability in the treatment response, which would therefore improve the precision of the treatment detection.[33] From a practical viewpoint, we considered using the baseline-uncorrected values to reduce the number of measurements needed for treatment classification. The baseline-uncorrected model would require only a single tapping assessment, whereas the baseline-corrected model would require two.

CROSS-VALIDATION We applied a nested k-fold cross-validation strategy to assess the performance and the generalizability of the composite biomarkers.[34] In nested cross-validation, the outer fold assesses the performance of the model, whereas the inner fold performs the model and hyper-parameter selection. In our study, the outer-fold step was repeated 100 times, with each iteration containing a different combination of training (80% of the data) and test sets (20%). Each outer training set was further split into an inner training (80% of the data) and validation sets (20%). The inner-fold step was repeated 50 times, and the best-performing inner model would be evaluated in the outer fold. The final results would be represented as the averaged and standard deviation of the models selected by each outer fold.[34] For the classification and regression models, we applied a group-shuffle split (same distribution of placebo and active treatments in each split) and a stratified-shuffle split (same distribution of MDS-UPDRS III scores in each split), respectively. To stratify the MDS-UPDRS III scores, we assigned each score to one of three binned ranges (eg, the baseline-corrected MDS-UPDRS III binned ranges were [-13, -8.76], [-8.76, -4.53], and [-4.53, 0.3]). Each outer fold had the same distribution of binned ranges. Stratification was not applied to the inner fold, as the small sample size would limit the number of samples available per bin. Within each inner fold, all features were standardized by subtracting the mean and scaling to the unit variance. To identify the features that were predictive of the outcomes, we identified features that were selected at least once by all outer-fold models.[34]

CLASSIFICATION OF ACTIVE OR PLACEBO TREATMENTS Classification models were trained to classify the active or placebo treatments. As we intended to predict the probability of treatment at all time points, we chose the last measurements to train the models. The MDS-UPDRS III classification model was trained on the 90-minute MDS-UPDRS III item scores.[14] The tapping classification models were trained on measurements taken immediately after the MDS-UPDRS III starting at 105 minutes. To identify the optimal classification model, we compared three classification models: support vector machines, logistic regression, and linear discriminant analysis (LDA). These classification models were selected as they are easy to implement and to interpret.[35-37] Previous studies have also used these algorithms to classify PD diagnosis or estimate MDS-UPDRS III.[38-41] Models were compared based on their mean accuracy, precision, and F1 scores.[40]

In addition, each model selected by the outer folds was used to predict the treatment at the other time points, with 20% of patients who were not used for training. This would allow researchers to identify at which time point treatment effects are detected. For each time point, the mean and standard deviation of the class probabilities were based on the predicted log-odd ratios from each fold. Additionally, these probabilities were used to estimate the repeatability and effect size. The repeatability was assessed by calculating the intraclass correlation coefficients (ICC) using the placebo results only. Using a random intercept model with the intercept and time point as fixed effects, the ICC was calculated by dividing the between-subject variance by the sum of the between-subject and within-subject variances. The effect size was calculated using all available data and a random intercept model with intercept, time point, treatment, and interaction between time point and treatment as fixed effects. In addition, the effect size was calculated as the contrast between the probabilities after treatment and the averaged baseline probabilities divided by the square root of the sum of the between-subject and within-subject variations.

ESTIMATION OF THE MDS-UPDRS III TOTAL SCORE To assess if the tapping composite biomarkers (baseline uncorrected and baseline corrected) could estimate the MDS-UPDRS III total score, linear regression with elastic-net regularization (optimized for α and the l1 ratio) was used to predict the MDS-UPDRS III total score at 90 minutes using the 105-minute tapping biomarkers. These two time points were compared, as it was previously shown that the IFT and TIFT showed significant and moderate-to-strong correlations with the MDS-UPDRS III.[14] Further, the 90- and 105-minute tapping tasks were equally as close to the 90-minute MDS-UPDRS III in timing and therefore we assumed would perform equally well.

To assess the performance of the models, we estimated the mean absolute error (MAE) of the outer-fold models. We evaluated the correlation between the predicted and true MDS-UPDRS III scores at all timepoints for each outer-fold model. Like the classification models, the MDS-UPDRS III scores were estimated at other time points with the 20% patients who were not used for training. Additionally, as for the classification models, those data were also used to estimate the repeatability and effect size.

Results

DATA COLLECTED

Twenty PD patients participated in this study. An overview of the demographic and disease characteristics of the patients was published previously ;[14] 14 patients were male, and their ages ranged from 48 to 70 years. Patients received one to four capsules of 100/25 mg L-dopa/carbidopa as they had a supramaximal morning levodopa equivalent dose (LED) ranging from 47 to 391 milligrams. The median MDS-UPDRS III score when using regular medication was 23 and 22 on their placebo and active treatment days, respectively.[14]

We analyzed 31 IMFT, 31 IFT, and 25 TIFT features. No features were excluded due to nonnormal distribution. Due to goniometer damage, we had missing data for 1 patient in the placebo condition and 2 patients in the active condition. As 6 patients had difficulties performing the IMFT, this led to missing data. However, the missing data were equally distributed across the treatment conditions and therefore deemed missing at random.

CLASSIFICATION OF PLACEBO AND ACTIVE TREATMENTS

We found that the LDA classifier consistently yielded the highest accuracy for all models (for both baseline uncorrected and baseline corrected); thus, we reported only the LDA results.

CLASSIFICATION OF TREATMENT EFFECTS The best-performing baseline-uncorrected composite biomarker, the IFT, yielded an accuracy, precision, F1 score, and large effect size of 68.50%, 70.23%, 68.93%, and 1.60 respectively (Table 1). The best-performing baseline-corrected composite biomarker, the IFT, achieved a higher average accuracy, precision, F1 score, and large effect size of 83.50%, 93.95%, 80.09%, and 2.58. Both models outperformed the MDS-UPDRS III classification models across all metrics. The IFT features that were mutually identified as important features for the baseline-uncorrected and baseline-corrected classification models were related to accuracy (e.g., spatial errors and the bivariate contour ellipse area), fatigue (e.g., velocity changes), and velocity (e.g., intertap intervals) (Figure 1).

CLASSIFICATION OF TREATMENT EFFECTS AT ALL TIME POINTS In Figure 2, the classification models were applied to all time points, showing the mean predicted probability of an active (>0.5) or placebo treatment (<0.5). In the baseline-corrected IFT, TIFT, and MDS-UPDRS III models, the mean predicted probability of a patient receiving a placebo treatment was consistently less than 0.5. In contrast, when active treatment was administered, the baseline-corrected IFT and MDS-UPDRS III model had a mean predicted probability above 0.5 from 60 minutes onward. The baseline-corrected IMFT and TIFT models crossed the 0.5 thresholds after 45 minutes. We found that the baseline-corrected IFT biomarker determined a large effect size (0.81) at 30 minutes, whereas the baseline- uncorrected IFT biomarker reached a large effect size of 0.84 at 60 minutes. The MDS-UPDRS III achieved a large effect size at 60 minutes (1.69 and 1.04 for baseline corrected and baseline uncorrected, respectively) (Figure S2). The MDS-UPDRS III demonstrated higher repeatability than the tapping tasks. Whereas the baseline-uncorrected MDS-UPDRS III biomarker obtained an excellent ICC, the IFT and TIFT both achieved good ICCs (0.78, 0.80) (42). However, the ICCs of the baseline-corrected MDS-UPDRS III and the IFT, IMFT, and TIFT biomarkers decreased to a moderate ICC range between 0.52 and 0.66.[42]

ESTIMATION OF MDS-UPDRS III

The mean MDS-UPDRS III total scores at 90 minutes for the placebo and active treatments were 33.5 and 22.0, respectively. When baseline-corrected, the mean MDS-UPDRS III scores for the placebo and active treatments were 0.3 and -13.0, respectively (Figure 3).

The best-performing baseline-uncorrected regression models were the TIFT and IFT composite biomarkers, which achieved the lowest average MAE of 10.31 and 10.36, respectively. In addition, the TIFT and IFT showed large effect sizes of 1.47 and 2.23, respectively, when estimating the MDS-UPDRS III. The best-performing baseline-corrected model was the IFT composite biomarker, which yielded the lowest average MAE of 7.87. For both the baseline-uncorrected and baseline-corrected models, the best-performing composite biomarkers outperformed that of the composite biomarkers of the three tasks. For the IFT features, the features that were mutually selected by both models were similar to that of the IFT classification features (Figure 2; Figure S1).

ESTIMATION OF MDS-UPDRS III AT ALL TIME POINTS The predicted and true MDS-UPDRS III scores were significantly correlated for the baseline-corrected and baseline-uncorrected models (Table 2). Once again, the best positive correlations were achieved by the TIFT baseline-uncorrected composite biomarker ($r = 0.58$, $P < 0.01$) and the IFT baseline-corrected composite biomarker ($r = 0.69$, $P < 0.01$). The greatest difference in the true MDS-UPDRS III scores between the placebo and active treatment interventions was at 90 minutes (Fig. 3). The tapping tasks achieved a moderate to good ICC (Table 2).

Discussion

DETECTION OF TREATMENT EFFECTS

The IFT biomarker (baseline corrected and baseline uncorrected) was, on average, more predictive of and more sensitive to treatment effects than the MDS- UPDRS III biomarker in terms of accuracy, precision, and clinical significance (as supported by the effect-size performances) (Table 1). This is significant as the ability to detect changes in aspects of motor function that may be missed by traditional assessments allows for a more sensitive measure of treatment efficacy. This can be valuable for detecting small and early changes in motor function that are indicative of a treatment response. The most important IFT features used to classify treatment effects are in concert with previous studies (Figure 1) that also identified that forearm movements relating to velocity, amplitude, and rhythm are sensitive to anti- parkinsonian drug effects.[10,15,43,44] We demonstrated that treatment effects were detected at 45 and 60 minutes for the TIFT and IFT composite biomarkers, respectively (Figure 2). This finding is notable as the mean onset of L-dopa/carbidopa action is about 50 minutes (45). This suggests that tapping tasks can detect the onset of oral L-dopa/carbidopa. The MDS-UPDRS III was not performed at 45 minutes, so it could not be determined whether the MDS-UPDRS III biomarker could detect treatment effects at 45 minutes. These findings further propound that the tapping tasks are practical and sensitive composite biomarkers for detecting motor response changes induced by anti- parkinsonian drugs (46). Further, the large effect sizes can potentially reduce sample size requirements and enhance power for future tapping task trials that assess treatment effects.

The performance of the classification models (except for the ICC) improved when the features were baseline corrected. Despite this, both models provide practical and clinical value. The baseline-uncorrected models required only a single measurement and represented the current motor function status. The baseline-corrected models require two measurements and represent the changes in motor function over time. The increased performance suggests that treatment response is dependent on the patient's tapping profile during their off state and adjusting for baseline removes variation in the L-dopa/carbidopa response.

ESTIMATION OF MDS-UPDRS III

We found that the baseline-corrected IFT biomarker, despite yielding the best performance among all the biomarkers, achieved a prediction error of approximately eight points and was significantly moderately correlated using the MDS-UPDRS III. The prediction error is comparable to existing sensor-based composite biomarkers used to estimate the MDS-UPDRS III. Studies using data sourced from an Axitvity AX3 (placed on the wrist and back or only the wrist) to estimate the gold standard achieved an MAE ranging from 4.29 to 6.29 points.[47,48] The tapping biomarkers predicted a smaller range of MDS-UPDRS III scores compared to that of the true MDS-UPDRS III scores (Figure 3). It is likely due to using only hand and forearm motor function assessments to predict the MDS-UPDRS III total scores, which includes motor assessments of other regions affected by PD, such as gait, facial expression, and speech.[4] As the correlations of the true and predicted MDS-UPDRS III scores were moderate (Table 2), the tapping biomarkers still showed concurrent validity with the gold standard. This suggests that the tapping biomarkers could provide clinicians with an understanding of the acute effects of drugs on motor fluctuations within a short monitoring period.

Despite the discrepancies between the true and predicted MDS-UPDRS III total scores, with the advancements in technology, it is not unusual for the performance of new clinical assessments to outperform the current gold standard. However, the discrepancy between the two assessments influences the accuracy estimates of the new clinical assessments, and as it would be interpreted as a prediction error.[49] Therefore, we argue that accurate estimation of the MDS-UPDRS III score is not essential for the adoption of the composite biomarker as a new complementary assessment for estimating symptom severity. Rather, the consequences resulting from the disagreement between the gold standard and the tapping composite biomarkers should be investigated.

FUTURE WORK

We demonstrated that the tapping composite biomarkers could detect the onset of oral L-dopa/carbidopa at 45 minutes. A follow-up study could investigate if the tapping composite biomarkers could detect an earlier onset of an even faster-acting antiparkinsonian drug, such as inhaled apomorphine that has an onset as early as 8 minutes.[8] This would further validate the sensitivity of the tapping composite biomarker to detect fast-acting dopaminergic drug effects.

Our sample size may limit the generalizability of this study's findings as a small sample size may not be representative of the broader population of patients with PD, making it difficult to generalize its results to a larger population.[50] This is particularly relevant for PD studies, where the disease can manifest in different ways and progress at different rates in different patients. To mitigate the effect of the small sample sizes, we employed cross-validation to bootstrap and validate the models against different groups of patients. We propose conducting a follow-up trial to implement the tapping tasks among more PD patients with more diverse MDS-UPDRS III profiles. The data collected from the trial can be used as an independent data set to assess the validity, reliability, and generalizability of our current methods. Although composite biomarkers have the advantage of capturing multiple aspects of motor function, the effects of individual components within the composite biomarker must be carefully examined to avoid misleading interpretations of the results. For example, a treatment that improves tapping speed but worsens tapping rhythm may result in an overall neutral effect, making it difficult to interpret the treatment's efficacy. Like other composite measures, such as the MDS-UPDRS III total score, it is crucial to examine the effects of each feature of the composite biomarker separately, as well as in conjunction with the overall composite score, to better understand the treatment's impact on finger motor function.

Conclusion

In conclusion, the IFT biomarker was more predictive of and sensitive to the detection of treatment effects than the MDS-UPDRS III biomarker; therefore, the tapping biomarkers appear to hold promise for evaluating the early and rapid effects of antiparkinsonian drugs. Moreover, the tapping task is easy to perform and can be done in clinical settings as well as at home by patients themselves, making it a practical and convenient method for monitoring disease progression and treatment response. Using tapping biomarkers, clinicians can obtain accurate and reliable data that can inform treatment decisions in real time.

TABLE 1 The mean and standard deviations of the accuracy, precision, F1 score, and effect size for each biomarker (at 90 minutes for MDS-UPDRS III and 105 minutes for the tapping task) are based on the 100 outer folds of the nested cross-validation

	Tasks	Accuracy	Precision	F1-SCORE	ICC	Effect-size
BASELINE-UNCORRECTED	IMFT	56.90% (±15.09%)	61.67% (±22.53%)	56.56% (±18.07%)	0.60 (± 0.25)	0.64 (± 0.57)
	IFT	**68.50%** (±12.56%)	**70.23%** (±16.31%)	**68.93%** (±14.9%)	0.78 (± 0.21)	**1.60** (± 0.82)
	TIFT	67.72% (±15.84%)	65.55% (±21.03%)	67.51% (±18.22%)	0.78 (± 0.22)	1.14 (± 0.80)
	All 3 Tasks	63.0% (±16.91%)	64.35% (±27.32%)	59.82% (±23.16%)	0.68 (± 0.29)	0.91 (± 0.68)
	MDS-UPDRS III item scores	63.75% (±11.25%)	61.20% (±10.9%)	68.90% (±11.52%)	**0.92** (± 0.10)	1.03 (± 0.60)
BASELINE-CORRECTED	IMFT	66.86% (±15.23%)	70.83% (±17.25%)	69.01% (±15.04%)	0.57 (± 0.17)	1.44 (± 0.98)
	IFT	**83.50%** (±10.74%)	**93.95%** (±11.25%)	**80.09%** (±14.92%)	0.53 (± 0.16)	**2.58** (± 0.90)
	TIFT	77.86% (±14.97%)	82.32% (±21.43%)	74.72% (±18.44%)	0.52 (± 0.17)	1.14 (± 0.80)
	All 3 Tasks	77.98% (±13.26%)	81.85% (±21.15%)	74.66% (±19.17%)	0.48 (± 0.18)	0.91 (± 0.61)
	MDS-UPDRS III item scores	75.75% (±14.45%)	79.95% (±17.64%)	73.93% (±16.42%)	**0.66** (± 0.11)	2.12 (± 1.25)

TABLE 2 Average correlation and ICC (95% CI) between the true and predicted MDS-UPDRS scores across all time points for the repeated nested cross-validation 100 outer-fold predictions.

	Tasks	Correlation coefficient (r)	p-value	ICC	Effect-size
BASELINE-UNCORRECTED	IMFT	0.10 [0.03, 0.16]	p<.05 [<.05, 0.05]	0.69 [0.65, 0.73]	0.67 [0.53, 0.81]
	IFT	0.52 [0.45, 0.59]	p<.01 [<.01, <.01]	0.80 [0.76, 0.83]	1.02 [0.91, 1.14]
	TIFT	**0.58 [0.53, 0.63]**	p<.05 [<.01, <.05]	0.78 [0.74, 0.82]	**1.47** [1.27, 1.67]
	All 3 Tasks	0.11 [0.04, 0.18]	p<.05 [<.05, 0.05]	0.66 [0.61, 0.71]	0.75 [0.62, 0.88]
BASELINE-CORRECTED	IMFT	0.34 [0.27, 0.40]	p<.05 [<.01, 0.06]	0.48 [0.44, 0.52]	1.10 [0.92, 1.28]
	IFT	**0.69 [0.65, 0.73]**	p<.001[<.001,<.005]	0.45 [0.42, 0.48]	**2.23** [2.01, 2.45]
	TIFT	0.65 [0.60, 0.69]	p<.001 [<.001, <.001]	0.50 [0.46, 0.54]	1.37 [1.20, 1.54]
	All 3 Tasks	0.56 [0.52, 0.61]	p<.05 [<.001, <.05]	0.43 [0.39, 0.47]	1.06 [0.91, 1.21]

SUPPLEMENTARY TABLE 1 Overview of features for the Alternate Index and Middle Finger Tapping (IMFT), Alternate Index Finger Tapping (IFT), Thumb-Index Finger Tapping (TIFT)(8)

Task	Endpoint (UNIT)	Acronyms
TIFT	Amplitude: Slope from linear regression of each tap's amplitude against time. (degrees and degrees/seconds)	Mean (TAM) Change (TAC)
TIFT	Angle frequency change: Change in peak tapping frequency over time (Hz/min) Angle change (degrees^2/s)	Frequency Mean (AFM) Frequency Change (AFC) Angle Mean (AAM), Angle Change (AAC)
IMFT, IFT	Bivariate contour ellipse angle (degree) Bivariate contour ellipse area (mm^2) BCEA represents the area of an ellipse which encompasses the fixation points	BCEA angle (BCT) BCEA area (BCA)
IMFT, IFT	Distance travelled between consecutive taps (centimetres)	Total (DTT) Average (DTA) Standard Deviation (DTS) Covariance (DTV) Change between first/last (DTD) Change between intervals (DTC)
IMFT, IFT, TIFT	Inter-Tap Interval: Time between two consecutive taps (milliseconds)	Average (ITA) Standard Deviation (ITS) Covariance (ITV) Change between (ITC) Change between first/last (ITD)
IMFT, IFT	Missed Taps: Total number of double/missed taps (DBLTT) Ratio good taps: total taps (DBLTR) (count)	Total number of double/missed taps (DBLTT) Ratio good taps: total taps (DBLTR)
IMFT, IFT	Number of Halts: Number of taps where the inter-tap interval is larger than 2 * ITM (count)	NOH
TIFT	Peak frequency area under the curve: The total power around the peak frequency in the power spectrum around the peak frequency (degrees^2)	Amplitude (FPA) Frequency (FPF) Area under the curve (FPP)
IMFT, IFT	Ratio good taps:total taps: Taps on the correct side (left/right) of the screen	TNT
IMFT, IFT	Spatial error: Sum of the Euclidean distances between each tap and the center of the target (millimeters)	Total (SET) Average (SEA) Standard Deviation (SES) Covariance (SEV) Change between (SED) Change between first/last (SEC)
IMFT, IFT, TIFT	Total number of taps	TNT
IMFT, IFT	Total taps inside and outside target	Taps within the target circle (TIT) Taps outside the target circle (TOT)

[continuation of Supplementary Table 1]

Task	Endpoint (UNIT)	Acronyms
IMFT, IFT	Mean of each finger tap's velocity (centimetres/minute)	Average (VEA) Standard Deviation (VES) Covariance (VEV) Change between first/last (VED) Change between intervals (VEC)
TIFT	Mean of each finger tap's velocity (degrees/second)^2	Mean (TVM) Change (TVC)
TIFT	Velocity Amplitude (degrees/second)^2	Velocity Amplitude Mean (VAM) Change (VAC)
TIFT	Velocity Closing: Average of the amplitude (i.e. angle) travelled per second for each tap when moving the index finger towards the thumb (closing); velocity extracted from the derivative of the amplitude (degrees/second)	Mean (CVM) Change (CVC)
TIFT	Velocity Frequency (Hz)	Mean (VFM) Change (VFC)
TIFT	Velocity Opening: Average of the amplitude (i.e. angle) travelled per second for each tap when moving the index finger away from the thumb (opening); velocity extracted from the derivative of the amplitude (degrees/s)	Mean (OVM) Change (OVC)

FIGURE 1 The average feature coefficients of the respective features selected by the LDA (linear discriminant analysis) classifier for each finger tapping task feature and the MDS-UPDRS III (Movement Disorder Society-Unified Parkinson's Disease Rating Scale, Part III) item score features (baseline-uncorrected and baseline-corrected models). The error bars represent the 95% confidence interval.

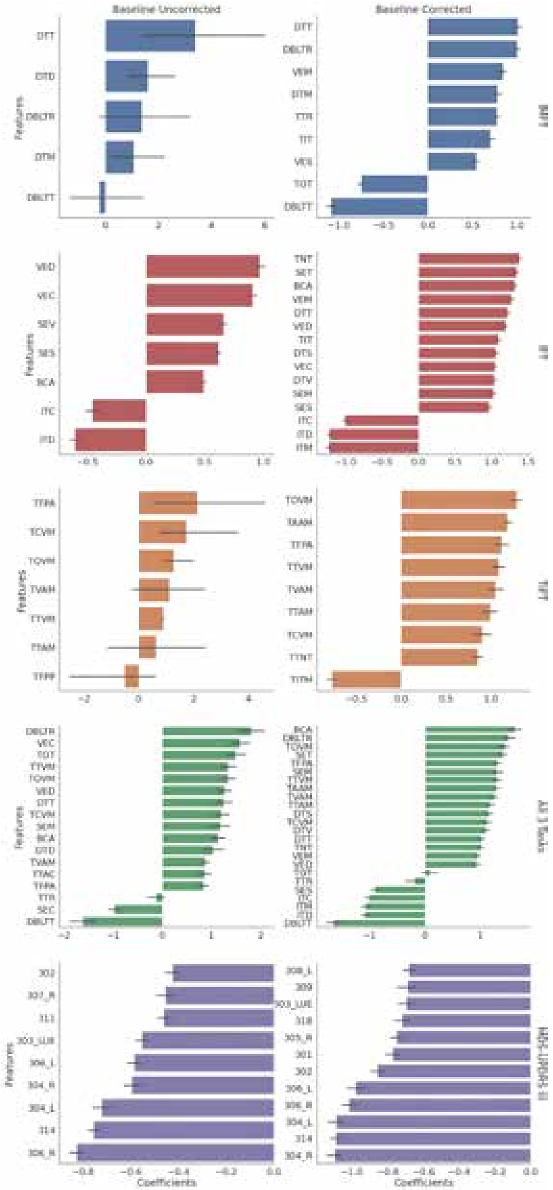


FIGURE 2 The mean predicted probability that active treatment was administered in the placebo (blue) and active (orange) treatment groups. The green dotted line represents the 0.5 decision boundary. The bands represent the 95% confidence interval.

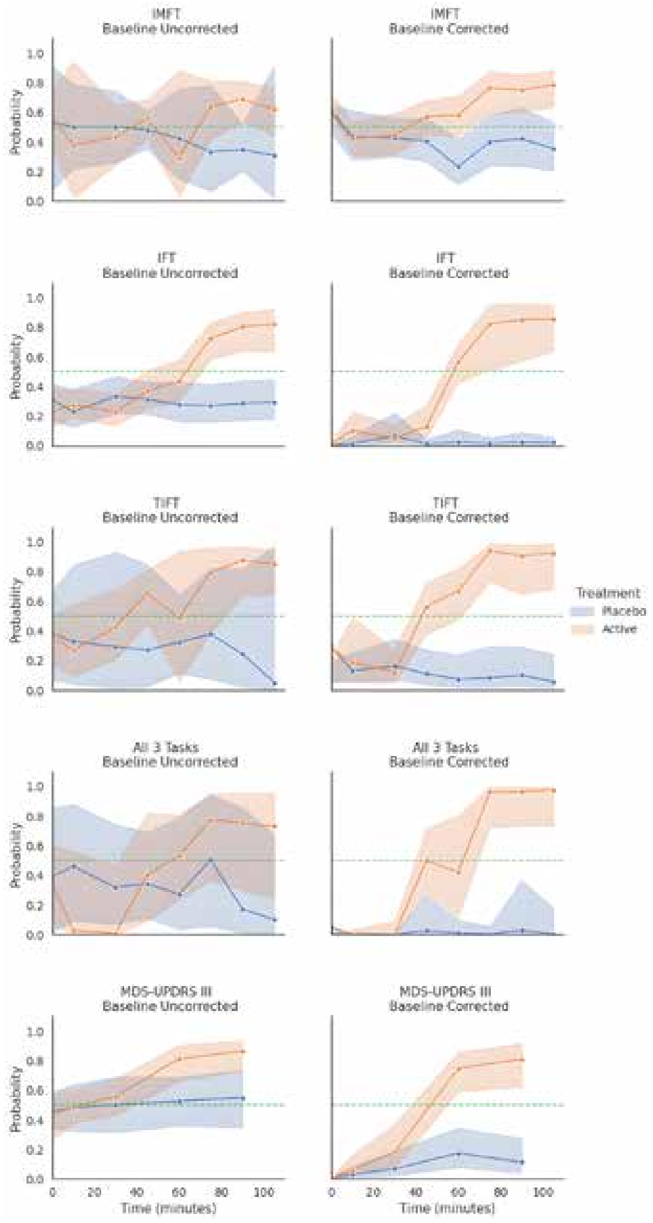
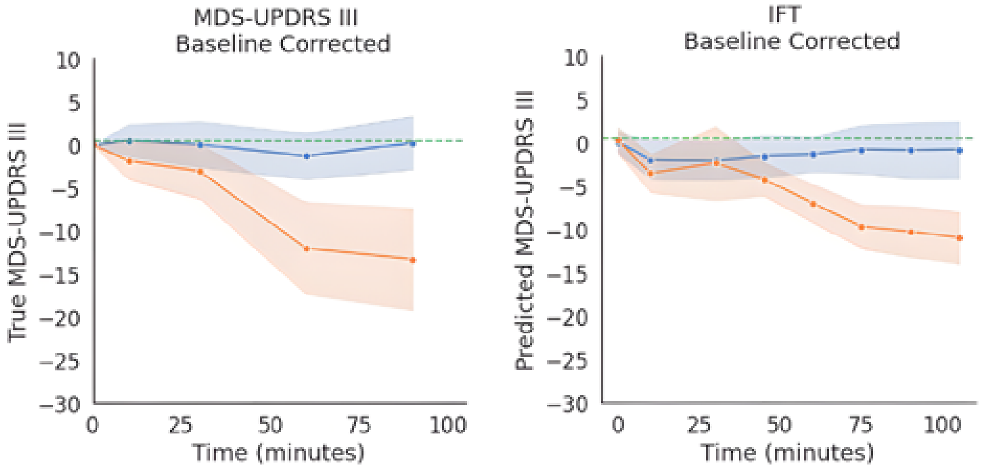
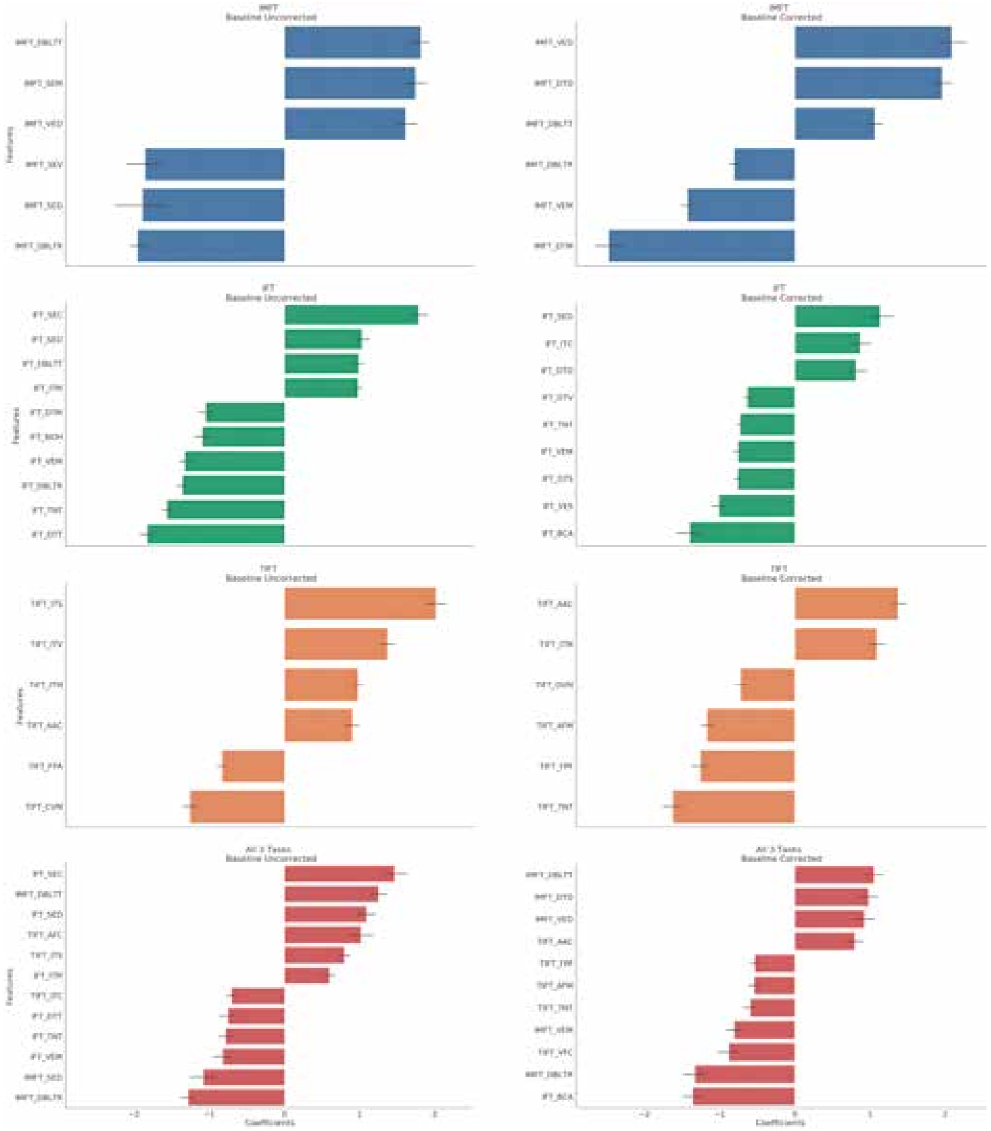


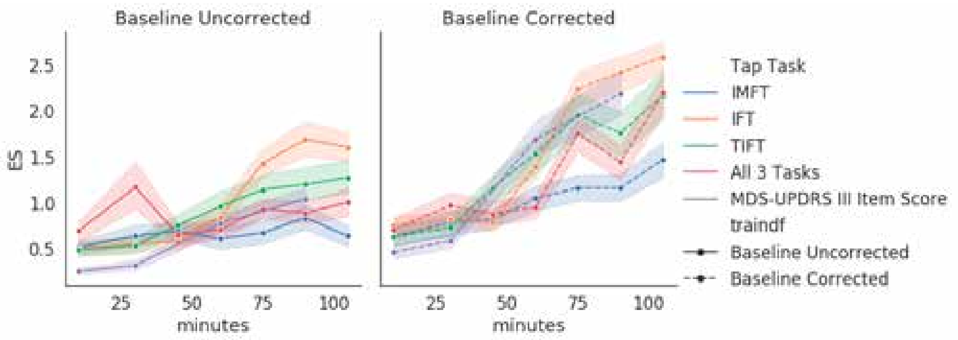
FIGURE 3 Average true and predicted MDS-UPDRS III (Movement Disorder Society-Unified Parkinson's Disease Rating Scale, Part III) scores with standard deviation from 0 to 105 minutes post dose for the placebo (blue) and active (orange) treatment interventions when baseline corrected.



SUPPLEMENTARY FIGURE 1 The average feature coefficients selected by the elastic-net linear regression models for each of the composite biomarkers under baseline-uncorrected and baseline-corrected conditions. The errors represent the 95% confidence intervals.



SUPPLEMENTARY FIGURE 2 Effect sizes of each of the tapping tasks and the Movement Disorder Society-Unified Parkinson's Disease Rating Scale, Part III, composite biomarkers at each time point.



REFERENCES

1 Davie CA. A review of Parkinson's disease. Br Med Bull 2008;86(1): 109–127. https://doi.org/10.1093/bmb/ldn013

2 Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008;79(4):368–376. https://doi.org/10.1136/jnnp.2007.131045

3 Regnault A, Boroojerdi B, Meunier J, Bani M, Morel T, Cano S. Does the MDS-UPDRS provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. J Neurol 2019;266:1927–1936. https://doi.org/10.1007/s00415-019-09348-3

4 Goetz CG 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 2008;23(15): 2129–2170. https://doi.org/10.1002/mds.22340

5 Martinez-Martin P, Rodriguez-Blazquez C, Alvarez-Sanchez M, et al. Expanded and independent validation of the Movement Disorder Society–unified Parkinson's disease rating scale (MDS-UPDRS). J Neurol 2013;260(1):228–236. https://doi.org/10.1007/s00415-012-6624-1

6 Ramsay N, Macleod AD, Alves G, et al. Validation of a UPDRS-/MDS-UPDRS-based definition of functional dependency for Parkinson's disease. Parkinsonism Relat Disord 2020;76:49–53. https://doi.org/10.1016/j.parkreldis.2020.05.034

7 Patel AB, Jimenez-Shahed J. Profile of inhaled levodopa and its potential in the treatment of Parkinson's disease: evidence to date. Neuropsychiatric Disease and Treatment 2018;14:2955–2964. https://doi.org/10.2147/NDT.S147633

8 Grosset KA, Malek N, Morgan F, Grosset DG. Inhaled apomorphine in patients with 'on-off' fluctuations: a randomized, double-blind, placebo-controlled, clinic and home based, parallel-group study. J Parkinsons Dis 2013;3(1):31–37. https://doi.org/10.3233/JPD-120142

9 Koop MM, Shivitz N, Brontë-Stewart H. Quantitative measures of fine motor, limb, and postural bradykinesia in very early stage, untreated Parkinson's disease. Mov Disord 2008;23(9):1262–1268. https://doi.org/10.1002/mds.22077

10 Makai-Bölöni S, Thijssen E, van Brummelen EMJJ, Groeneveld GJ, Doll RJ. Touchscreen-based finger tapping: repeatability and configuration effects on tapping performance. PLoS One 2021;16(12): e0260783. https://doi.org/10.1371/journal.pone.0260783

11 Nalçaci E, Kalaycioglu C, Çiçek M, Genç Y. The relationship between handedness and fine motor performance. Cortex 2001; 37(4):493–500. https://doi.org/10.1016/S0010-9452(08)70589-6

12 Taylor Tavares AL, Jefferis GSXE, Koop M, Hill BC, Hastie T, Heit G, Bronte-Stewart HM. Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation. Mov Disord 2005; 20(10):1286–1298. https://doi.org/10.1002/mds.20556

13 Fukawa K, Okuno R, Yokoe M, Sakoda S, Akazawa K. Estimation of UPDRS finger tapping score by using artificial neural network for quantitative diagnosis of Parkinson's disease. Proceedings of the IEEE/EMBS Region 8 International Conference on Information Technology Applications in Biomedicine, ITAB; IEEE, New York City; 2007:259–260. https://doi.org/10.1109/ITAB.2007.4407396.

14 Thijssen E, Makai-Bölöni S, van Brummelen E, den Heijer J, Yavuz Y, Doll RJ, Groeneveld GJ. A placebo-controlled study to assess the sensitivity of finger tapping to medication effects in PD. Mov Disord Clin Pract 2022;9:1074–1084. https://doi.org/10.1002/mdc3.13563

15 Espay AJ, Giuffrida JP, Chen R, et al. Differential response of speed, amplitude, and rhythm to dopaminergic medications in Parkinson's disease. Mov Disord 2011;26(14):2504–2508. https://doi.org/10.1002/mds.23893

16 Hasan H, Burrows M, Athauda DS, et al. The Bradykinesia Akinesia Incoordination (BRAIN) tap test: capturing the sequence effect. Mov Disord Clin Pract 2019;6(6):462–469. https://doi.org/10.1002/mdc3.12798

17 Wissel BD, Mitsi G, Dwivedi AK, et al. Tablet-based application for objective measurement of motor fluctuations in Parkinson disease. Digit Biomark 2018;1(2):126–135. https://doi.org/10.1159/000485468

18 Lipp MM, Batycky R, Moore J, Leinonen M, Freed MI. Preclinical and clinical assessment of inhaled levodopa for OFF episodes in Parkinson's disease. Sci Transl Med 2016;8(360):360ra136–360ra136. https://doi.org/10.1126/scitranslmed.aad8858

19 Arora S et al. Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD. Neurology 2018;91(16): E1528–E1538. https://doi.org/10.1212/WNL.0000000000006366

20 Kimber TE, Tsai CS, Semmler J, Brophy BP, Thompson PD. Voluntary movement after pallidotomy in severe Parkinson's disease. Brain 1999;122(5):895–906. https://doi.org/10.1093/brain/122.5.895

21 Yokoe M, Okuno R, Hamasaki T, Kurachi Y, Akazawa K, Sakoda S. Opening velocity, a novel parameter, for finger tapping test in patients with Parkinson's disease. Parkinsonism Relat Disord 2009;15(6):440–444. https://doi.org/10.1016/j.parkreldis.2008.11.003

22 Espay AJ, Hausdorff JM, Sanchez-Ferro ´A, et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord 2019;34(5):657–663. https://doi.org/10.1002/mds.27671

23 Nisenzon AN, Robinson ME, Bowers D, Banou E, Malaty I, Okun MS. Measurement of patient-centered outcomes in Parkinson's disease: what do patients really want from their treatment? Parkinsonism Relat Disord 2011;17(2):89–94. https://doi.org/10.1016/j.parkreldis.2010.09.005

24 Sikap P et al. Perancangan Prototipe Sistem Pemesanan Makanan dan Minuman Menggunakan Mobile Device. Indonesia Journal on Networking and Security 2015;1(2):1–10. https://doi.org/10.1145/242224.242229

25 Zhan A, Mohan S, Tarolli C, et al. Using smartphones and machine learning to quantify Parkinson disease severity the mobile Parkinson disease score. JAMA Neurol 2018;75(7):876–880. https://doi.org/10.1001/jamaneurol.2018.0809

26 Mei J, Desrosiers C, Frasnelli J. Machine learning for the diagnosis of Parkinson's disease: a review of literature. Frontiers in Aging Neuroscience 2021;13:184. https://doi.org/10.3389/fnagi.2021.633752

27 Yang N, Liu DF, Liu T, et al. Automatic detection pipeline for accessing the motor severity of Parkinson's disease in finger tapping and postural stability. IEEE Access 2022;10:66961–66973. https://doi.org/10.1109/access.2022.3183232

28 Kalia LV, Lang AE. Parkinson's disease. The Lancet 2015;386(9996): 896–912. https://doi.org/10.1016/S0140-6736(14)61393-3

29 Tomlinson CL, Stowe R, Patel S, Rick C, Gray R, Clarke CE. Systematic review of levodopa dose equivalency reporting in Parkinson's disease. Mov Disord 2010;25(15):2649–2653. https://doi.org/10.1002/mds.23429

30 Biometrics Ltd. Twin-Axis goniometers for dynamic joint movement analysis; 2020.

31 Van Rossum G, Drake FL Jr. Python 3 Reference Manual, Version 3.7.3. Scotts Valley, CA: CreateSpace; 2009.

32 Pedregosa F. Scikit-learn: machine learning in {P}ython. Journal of Machine Learning Research 2011;12:2825–2830.

33 Kaiser L. Adjusting for baseline: change or percentage change? Stat Med 1989;8(10):1183–1190. https://doi.org/10.1002/sim.4780081002

34 Parvandeh S, Yeh HW, Paulus MP, McKinney BA. Consensus features nested cross-validation. Bioinformatics 2020;36(10): 3093–3098. https://doi.org/10.1093/bioinformatics/btaa046

35 Navia-Vazquez A, Parrado-Hernandez E. Support vector machine interpretation. Neurocomputing 2006;69(13–15):1754–1759. https://doi.org/10.1016/j.neucom.2005.12.118

36 Deng Y, Liu X, Xin C, Jia W. An interpretable classifier with linear discriminant analysis based on AFS theory. 2019 Chinese Control Conference (CCC). IEEE, New York City; 2019:7583–7588. https://doi.org/10.23919/ChiCC.2019.8866096.

37 Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. WIREs Data Mining and Knowledge Discovery 2020;10(5):e1379. https://doi.org/10.1002/widm.1379

38 Moon S, Song HJ, Sharma VD, Lyons KE, Pahwa R, Akinwuntan AE, Devos H. Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach. J Neuroeng Rehabil 2020;17(1):125. https://doi.org/10.1186/s12984-020-00756-5

39 Wu Y, Krishnan S. Statistical analysis of gait rhythm in patients with Parkinson's disease. IEEE Trans Neural Syst Rehabil Eng 2010;18(2):150–158. https://doi.org/10.1109/TNSRE.2009.2033062

40 Geetha R, Sivagami G. Parkinson Disease Classification using Data Mining Algorithms; 2011.

41 Yadav G, Kumar Y, Sahoo G. Predication of Parkinson's disease using data mining methods: a comparative analysis of tree, statistical and support vector machine classifiers. 2012 National Conference on Computing and Communication Systems. IEEE, New York City; 2012:1–8. https://doi.org/10.1109/NCCCS.2012.6413034

42 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016; 15(2):155–163. https://doi.org/10.1016/j.jcm.2016.02.012

43 Trager MH, Velisar A, Koop MM, Shreve L, Quinn E, Bronte- Stewart H. Arrhythmokinesis is evident during unimanual not bimanual finger tapping in Parkinson's disease. J Clin Mov Disord 2015;2(1):8. https://doi.org/10.1186/s40734-015-0019-2

44 Giovannoni G, Van Schalkwyk J, Fritz VU, Lees AJ. Bradykinesia akinesia incoordination test (BRAIN TEST): an objective computerized assessment of upper limb motor function. J Neurol Neurosurg Psychiatry 1999;67(5):624–629. https://doi.org/10.1136/jnnp.67.5.624

45 Hauser RA, Ellenbogen A, Khanna S, Gupta S, Modi NB. Onset and duration of effect of extended-release carbidopa-levodopa in advanced Parkinson's disease. Neuropsychiatr Dis Treat 2018;14:839–845. https://doi.org/10.2147/NDT.S153321

46 Contin M, Riva R, Martinelli P, Albani F, Avoni P, Baruzzi A. Levodopa therapy monitoring in patients with Parkinson disease: akinetic-dynamic approach. Ther Drug Monit 2001;23(6):621–629. https://doi.org/10.1097/00007691-200112000-00005

47 Lobo V, Branco D, Guerreiro T, Bouça-Machado R, Ferreira J. Machine-learning models for MDS-UPDRS III prediction: a comparative study of features, models, and data sources. Information Society 2022.

48 Ur Rehman RZ, Rochester L, Yarnall AJ, Del Din S. Predicting the progression of Parkinson's disease MDS-UPDRS-III motor severity score from gait data using deep learning. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBS, Institute of Electrical and Electronics Engineers Inc., New York City; 2021:249–252. https://doi.org/10.1109/EMBC46164.2021.9630769.

49 Walsh T. Fuzzy gold standards: approaches to handling an imperfect reference standard. J Dent 2018;74:S47–S49. https://doi.org/10.1016/j.jdent.2018.04.022

50 Berisha V, Krantsevich C, Hahn PR, Hahn S, Dasarathy G, Turaga P, Liss J. Digital medicine and the curse of dimensionality. npj Digital Medicine 2018;4(1):1–8. https://doi.org/10.1038/s41746-021-00521-5