Exploring test retest reliability and longitudinal stability of digital biomarkers for Parkinson's disease in the m-Power dataset

Authors
Mehran Sahandi Far[1,2], Simon B. Eickhoff [1,2], María Goñi[1,2*], Juergen Dukart[1,2*]

Affiliations

[1] *Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany*

[2] *Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany*

**Contributed equally*

Word count (abstract/main text/methods): 231 / 1825 / 1160

To whom correspondence should be addressed:
Juergen Dukart, PhD
Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7),
Research Centre Jülich, Jülich, Germany
Email: juergen.dukart@gmail.com

Abstract

Digital biomarkers (DB) as captured using sensors embedded in modern smart devices are a promising technology for home-based symptom monitoring in Parkinson's disease (PD). Despite extensive application in recent studies test-retest reliability and longitudinal stability of DB has not been well addressed in this context. We utilized the large-scale m-Power dataset to establish the test-retest reliability and longitudinal stability of gait, balance, voice and tapping tasks in an unsupervised and self-administered daily life setting in PD patients and healthy volunteers. Intraclass Correlation Coefficients (ICC) were computed to estimate the test-retest reliability of features that also differentiate between PD and healthy volunteers. In addition, we tested for longitudinal stability of DB measures in PD and HC as well as for their sensitivity to PD medication effects. Among the features differing between PD and HC, only few tapping and voice features had good to excellent test-retest reliabilities and medium to large effect sizes. All other features performed poorly in this respect. Only few features were sensitive to medication effects. The longitudinal analyses revealed significant alterations over time across a variety of features and in particular for the tapping task. These results indicate the need for further development of more standardized, sensitive and reliable DB for application in self-administered remote studies in PD patients. Motivational, learning and other confounds may cause a variation in performance that needs to be considered in DB longitudinal applications.

Introduction

Parkinson's disease (PD) is primarily characterized by motor symptoms including tremor at rest, rigidity, akinesia, and postural instability [1]. Although standard in-clinic assessments such as the Unified Parkinson's Disease Rating Scale (UPDRS) are popular, they are influenced by inter-rater variability by relying on self-reporting by patients and care givers or clinician judgement [2]. In addition, they are costly and limited with respect to observation frequency.

The emergence of new technologies has led to a variety of sensors (i.e. acceleration, gyroscope, GPS, etc.) embedded in smart devices of daily use (i.e., smartphone, smartwatch). Such sensor data alongside with other digital information recorded passively or when executing prespecified tasks may provide valuable insight into health-related information. Such applications are now commonly referred to as digital biomarkers (DB) [3]. DB being collected frequently over a long period of time can provide an objective, ecologically valid and more detailed understanding of the inter- and intra-individual variability in disease manifestation in daily life.

Numerous DB have been proposed for PD diagnosis as well as to establish a link with clinical rating scales such as UPDRS and to quantify disease severity or intervention effect [4–8]. Despite these various proof of concept studies, many technical challenges with respect to DB deployment remain unaddressed. DB measures are prone to large variation caused by technical and procedural differences including but not limited to placement/orientation, recording frequency of the devices and environmental and individual variation, i.e. due to motivation, medication or other aspects [9–11]. Other factors such as the effect of users' familiarity with technology and the impact of learning on the performance of measured DB in remote PD assessment are other important sources of variation that have not been addressed so far. All of these factors may limit the sensitivity and reliability of DB measurements for any of the above PD clinical applications. DB longitudinal variation is therefore an important attribute that should be quantified and addressed. The reliability of DB assessment has been broadly studied for gait, balance, voice and tapping data [12–17]. However, the existing studies typically focused on a single or few aspects of PD and most of them established the test-retest reliability in a standardized clinical setting limiting the translatability of their findings to at-home applications. Among studies that evaluated DB assessments for remote monitoring of PD, only one reported the test-retest reliability [5]. No

PD studies systematically evaluated the test-retest reliability and longitudinal sensitivity of DB in a fully unsupervised and self-administered PD longitudinal setting.

Here aimed to address these open questions on the performance of DB measures in PD when collected in a self-administered setting in daily life. For this, we first performed a comprehensive literature search identifying 773 DB features reported in previous studies to cover PD-related alterations in gait characteristics, tremor, postural instability, voice and finger dexterity. We evaluated the longitudinal stability and test re-test reliability of these features as collected using four commonly applied PD tasks (gait, balance, voice and tapping) in daily life in a large cohort of self-reported PD patients and healthy controls (HC), the m-Power study [18]. In addition, we evaluated their sensitivity to learning and medication effects.

[Table 1]

Results

[Figure 1]

Differentiation between PD and HC

First, we aimed to restrict the test-retest reliability analyses of the initial 773 features to those which significantly differ between PD (N=610 to 970 depending on the task, Table 1) and HC (N=807 to 1674). For this, we performed group comparisons for all computed features for gait, balance, voice and tapping tasks. For gait 66 out of 423, for balance 59 out of 183, for voice 62 out of 124 and for tapping 25 out of 43 features differed significantly ($p<0.05$) between PD and HC at baseline with small (gait and balance) to medium effect sizes for gait, balance and voice and small to large effect sizes for the tapping task (Figure 1 and Supplement 2).

[Figure 2]

Test-retest reliability

Next, we identified separately for PD and HC the top 10 features with highest median test retest reliability (as measured using intraclass correlation coefficients – ICC) across different time points (one hour, one day, one week or one month apart) and repetitions (all subjects with 5 repetitions of the task). This procedure resulted in 12 to 15 features (including shared ones) being selected for each task (Figure 2, S1 and S2). ICC analyses revealed poor to

good test-retest reliability for these most reliable features from the gait and balance tasks and good to excellent reliability for features from voice and tapping tasks (Figure 2). The average ICC across the best performing features selected from different repetitions was lower at the fifth repetition compared to the first, it dropped from .11 to .09 for the gait task, from .21 to .13 for the balance task, from .39 to .24 for the voice task and from .3 to .23 for the tapping task. The average ICC across the best performing features selected from different time points was also lower at one-month compared to one-hour, it dropped from .13 to .07 for the gait, from .2 to .12 for the balance, from .33 to .26 for the voice and from .32 to .19 for the tapping task.

[Figure 3]

Repetition effects

Next, we evaluated the longitudinal stability of these most reliable features. Using repeated measures analyses of variance, we tested for main effects of diagnosis, repetition (1st, 2nd, 3rd, $4^{th,}$ and 5th) and their interaction (Figure 3,4, Table S5). A significant main effect of diagnosis across all time points was observed for 6 out of 15 gait features, 11 out of 15 balance features, 8 out of 12 voice features and 11 out of 12 tapping features. A significant effect of repetition was found for 8 out of 15 gait features, 8 out of 15 balance features, 4 out of 12 voice features 10 out of 12 tapping features. A significant diagnosis-by-repetition interaction effect was identified for 3 out of 15 gait features, 0 out of 15 balance features, 3 out of 12 voice features, and 9 out of 12 tapping features.

[Figure 4]

Medication effects

Lastly, we tested which of the most reliable features identified above also display sensitivity to PD medication. For this we compared the conditions reported by the patients as being before PD medication, after PD medication or at best. A significant effect of PD medication was only observed for 2 out of 15 gait features, 1 out of 15 balance features, 2 out of 12 voice features and 1 out of 12 tapping features (Figure S3, Table S5-medication column).

Discussion

Here we assessed the longitudinal test-retest reliability and stability of DB measures related to gait, balance voice and finger dexterity impairments in PD. We found a wide range of test-retest reliabilities across tasks and features ranging from poor to excellent with highest reliabilities observed for voice followed by the tapping task. Only very few features had medium to large effects sizes for differentiation between PD and HC. For all tasks, a substantial percentage of features displayed significant longitudinal alterations.

Overall, tapping and voice tasks revealed a better performance compared to gait and balance tasks with respect to test-retest reliability and observed effect sizes. Both, balance and gait tasks displayed consistently poor test-retest reliabilities as well as low effect sizes for differentiation between PD and HC questioning their usability for home-based applications. In contrast, best performing voice features displayed fair to excellent test-retest reliabilities across repetitions but also over weeks and months. Most features showed a drop in test-retest reliability with longer periods of time. This may potentially reflect a consequence of the repetition effects and the group-by-repetition interactions observed in the analyses of variance for a substantial proportion of the features.

Despite significant difference at baseline, several features did not differentiate PD and HC when using data from all time points. This effect became most pronounced for the gait task, likely due to its poor test-retest reliability performance. Differential learning, variation in motivation, medication, reduced adherence to task instructions, and other physical and environmental parameters may contribute to this loss of differentiation [29,11]. Whilst a clear differentiation of motivation vs learning effects on the often abstract DB features is difficult in an observational study design, a possible way to provide inference on this issue is to compare the direction of alterations in PD and HC. Assuming that alterations in PD relative to HC reflect impairment, movement of a feature state towards PD is likely to reflect worsening either due to reduced motivation, disease progression or other similar factors. In contrast, movements towards HC is likely to reflect improvement and is therewith compatible with a learning effect. We find a mixture of both effects for most tasks suggesting the presence of both aspects in DB longitudinal data. These observations are also in line with previous studies showing that training may reduce motor impairment in PD [19–21]. In particular, for the tapping task the difference between PD and HC disappears for several features which is primarily due to a shift in performance in HC. These findings may point to a differential change in motivation across groups. Whilst differential learning has been previously reported [19,22,23], the differential change in motivation is an important novel aspect

to consider when comparing DB measures between PD patients and HC. Understanding the sources leading to this variability of DB measures over time is a vital and open-question which needs to be systematically addressed to enable their application for specific clinical questions.

Most patients with PD take dopaminergic medication to alleviate their motor functions. However, the responsiveness to PD medication highly varies between patients. Besides good reliability and the ability to differentiate PD and HC, another important and desired quality of an effective DB is therefore to monitor PD medication response. Among the most reliable features from each task, only very few displayed significant but weak sensitivity to different medication conditions. One possible reason for this poor performance of DB measures in our study as compared to some previous reports [24], might be the self-reported nature of the medication status in the m-Power dataset likely introducing some noise variation (i.e. different drugs and differences in time after administration). Nonetheless, our findings point to the need for further optimization of DB measures to increase their sensitivity to PD medication effects.

The self-administered design of the m-Power dataset is also the major limitation of our study. In such an uncontrolled setting, accuracy in reporting the diagnosis and demographics, defining the medication status but also correct understanding and compliance with the instructions may all have introduced variation into the study measures. The reported ballpark estimates for test-retest reliability and ability of the respective measures to differentiate between PD and HC therefore need to be carefully considered when interpreting our results. Nonetheless, our findings clearly demonstrate the need for further optimization of DB tasks as well for introducing careful monitoring and quality control procedures to enable integration of DB measures into clinically relevant applications.

Methods

Study cohort

We utilized smartphone-based data from a large-scale longitudinal observational cohort collected in the first six month of m-Power study [18]. The enrolment was open to adult participants who own an iPhone, are living in the US and are comfortable with English to read the instructions in the app. Participants were asked to download the application and complete a one-time demographic survey during registration. Demographic data includes but not limited to age, sex, health history, and previous PD clinical diagnosis. They also were

asked to fill out a survey with selected questions from the movement disorder society's UPDRS, as well as, PDQ-8. All the participants were suggested to complete each task (walking, tapping, voice, and memory) up to three times a day for up to six months. In addition, self-reported PD patients were asked to complete the task before medication, after medication and at another time when they are feeling at their best.

Data Pre-processing

The m-Power dataset is assessed outside of clinical environment with severely limited quality control and supervision. All information including the health history, diseases diagnosis, duration, treatment and survey outcomes are self-reported. To address these, we excluded participants, who did not specify their age, sex, or information on professional diagnosis (if they belong to the PD or HC group). The participants are assigned to PD or HC group according to their answer to the question "Have you been diagnosed by a medical professional with Parkinson disease?". There was a significant difference in the age and sex distribution between HC and PD subjects, particularly age slanted toward younger and male individuals in HC. To reduce the impact of age, we restricted the age range for our analysis to between 35 and 75 years. The resulting number of available assessments per subject is displayed in Table 1.

Feature extraction

To identify features that are commonly used for the walking, voice and tapping tasks for PD applications, we performed a comprehensive literature search in PubMed with the following terms ((Parkinson's disease) AND (walking OR gait OR balance OR voice OR tapping) AND (wearables OR smartphones)). Based on this search, we identified an overall of 773 features related to gait (N=423), balance (N=183), finger dexterity (N=43), and speech impairment (N=124). All of these features were computed for the m-Power study. A detailed explanation of the extracted features including the respective references is provided in Table S1-S4. For features sharing the same variance (high pair-wise correlation: Spearman rho>.95), only one of the features was selected randomly for further analyses to reduce the amount of redundant information for each task.

Gait and balance: Impairments in gait speed, stride length, and stride-time-variability are common changes that are linked to PD [25–28]. Instability in postural balance is also considered as one of the well-reported characteristics associated with PD [14,29–31]. Both were

assessed by a walking task. The gait part consisted of 20 steps walking in a straight line, followed by the balance part of a 30 seconds stay still period. Given a heterogeneity of gait signal lengths across subjects, we used a fixed length signal of 10 seconds and selected data from participants who met this criterion which resulted in 28150 records from 1417 unique participants. In addition to the accelerometer signals (x, y and z), their average, the step series, position along the three axes by double integration, and velocity and acceleration along the path were used for feature extraction [32,33] (Table S1). For balance, we used a 15 seconds time-window trimming the first five and the last 10 seconds of the 30 second records to reduce the noise due to the between-task transition period, resulting in 29050 records from 1435 unique participants. Feature extraction covered signals related to tremor acceleration predicted to fall in the 4-7 Hz band and postural acceleration (non-tremor) falling to the 0-3.5 Hz band [34] (Table S2).

Voice: PD may also affect breathing and results in alterations in speech and voice. Reduced volume, hoarse quality, and vocal tremor are commonly reported for PD using voice analysis [15,35,36]. In this task, participants were saying "aaaah" for about ten seconds. For voice, 49676 records were selected belonging to 2184 unique participants. Voice features were computed from fundamental frequency, amplitude and period signals trimming the first and the last two seconds of the ten second interval (Table S3).

Tapping: Impairment in finger dexterity is another symptom associated with PD [37,38]. In the m-Power study, participants were asked to tap as fast as possible for 20 seconds with the index and middle fingers on the screen of their phone (positioned on a flat surface). Screen pixel coordinate (x, y) and timestamp of taped points plus acceleration sensor data were collected for this task. Overall, 55894 recordings were selected belonging to 2644 unique participants. Features were computed based on the inter-tapping distance and interval (Table S4).

Statistical analysis

For features to be considered usable for biomarker purposes in longitudinal studies several criteria are important including among other sensitivity to disease symptoms, good test-retest reliability and robustness against the effects of learning and other longitudinal confounds. To address these criteria, we adopted a step-wise statistical procedure.

As DB measures are frequently not normally distributed, Mann-Whitney U test were used to identify all features that significantly differ between PD and HC at the first administration (baseline) ($p<0.05$). Effect sizes Cohen's d were computed for these features to provide an estimate of the magnitude of differentiation between PD and HC.

Next, Intraclass Correlation Coefficients (ICC, type 1-1) were used to determine the test-retest reliability of features showing a significant differentiation between PD and HC. ICC Values of 0.0–0.40 were considered as poor, 0.40–0.59 as fair, 0.60–0.74 as good, and 0.75–1.00 to be excellent [39]. To assess the reliability of each feature, ICCs were computed for different time points vs baseline (one hour (0-6 hours), one day (calendric day), one week (7 calendric days) or one month apart (30 calendric days)), as well as for different repeats versus baseline (baseline versus 2nd, 3rd, 4^{th,} and 5th repeat). We then selected the top 10 features with the highest median ICCs for each group (PD, HC) and tested for their longitudinal stability over time. For this, we computed repeated measure analyses of variance (rm-ANOVA) using a mixed factorial design with a between-subject factor diagnosis and a within-subject factor repetition (1st, 2nd, 3rd, 4^{th,} and 5th) including their interaction. Patients who had at least four repetitions after baseline (463 for gait, 597 for balance, 1085 for voice and 1333 for tapping) were included in these analyses. Lastly, we assessed the impact of PD medication by computing rm-ANOVAs in the PD group with the within-subject factor medication (i.e. before, after, and at best). Patients who had at least one marked task for each of the three PD medication conditions (i.e. before, after, and at best) were included in treatment effect analysis (188 for gait, 189 for balance, 280 for voice and 338 for tapping).

Data availability

The m-Power dataset used for this article is available upon registration from Synapse at: https://www.synapse.org/#!Synapse:syn4993293.

Acknowledgements

This study was supported by the Human Brain Project, funded from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2).

The data used in this study were contributed by users of the Parkinson m-Power mobile application as part of the m-Power study developed by Sage Bionetworks and described in Synapse [doi:10.7303/syn4993293].

Author contribution

MSF performed analyses and wrote manuscript. MG performed feature extraction. MSF, JD and MG contributed to study design and writing the manuscript. SBE and JD designed the overall study and contributed to interpretation of the results. All authors reviewed and commented on the manuscript.

Competing interest

JD is a former employee and received consultancy fees on another topic from F. Hoffmann-La Roche AG. All authors report no conflicts of interest with respect to the work presented in this study.

References

1. Jankovic, J. Parkinson's disease: clinical features and diagnosis. *J. Neurol. Neurosurg. Psychiatry* **79,** 368–376 (2008).
2. Prince, J., Arora, S. & de Vos, M. Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. *Physiol Meas* **39,** 044005 (2018).
3. Insel, T. R. Digital phenotyping: technology for a new science of behavior. *JAMA* **318,** 1215–1216 (2017).
4. Mahadevan, N. *et al.* Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *npj Digital Med.* **3,** 5 (2020).
5. Lipsmeier, F. *et al.* Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Mov. Disord.* **33,** 1287–1297 (2018).
6. Shah, V. V. *et al.* Digital biomarkers of mobility in parkinson's disease during daily living. *J. Parkinsons. Dis.* **10,** 1099–1111 (2020).
7. Schlachetzki, J. C. M. *et al.* Wearable sensors objectively measure gait parameters in Parkinson's disease. *PLoS One* **12,** e0183989 (2017).
8. Tracy, J. M., Özkanca, Y., Atkins, D. C. & Hosseini Ghomi, R. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. *J. Biomed. Inform.* **104,** 103362 (2020).
9. Espay, A. J. *et al.* Technology in Parkinson's disease: Challenges and opportunities. *Mov. Disord.* **31,** 1272–1282 (2016).
10. Moore, S. T. *et al.* Autonomous identification of freezing of gait in Parkinson's disease from lower-body segmental accelerometry. *J Neuroeng Rehabil* **10,** 19 (2013).
11. Fisher, J. M., Hammerla, N. Y., Rochester, L., Andras, P. & Walker, R. W. Body-Worn Sensors in Parkinson's Disease: Evaluating Their Acceptability to Patients. *Telemed J E Health* **22,** 63–69 (2016).
12. Rahlf, A. L., Petersen, E., Rehwinkel, D., Zech, A. & Hamacher, D. Validity and Reliability of an Inertial Sensor-Based Knee Proprioception Test in Younger vs. Older Adults. *Front. Sports Act. Living* **1,** (2019).
13. Orlowski, K. *et al.* Examination of the reliability of an inertial sensor-based gait analysis

system. *Biomed Tech (Berl)* **62,** 615–622 (2017).
14. Hasegawa, N. *et al.* How to Select Balance Measures Sensitive to Parkinson's Disease from Body-Worn Inertial Sensors-Separating the Trees from the Forest. *Sensors (Basel)* **19,** (2019).
15. Skodda, S., Grönheit, W., Mancinelli, N. & Schlegel, U. Progression of voice and speech impairment in the course of Parkinson's disease: a longitudinal study. *Parkinsons. Dis.* **2013,** 389195 (2013).
16. Aghanavesi, S., Nyholm, D., Senek, M., Bergquist, F. & Memedi, M. A smartphone-based system to quantify dexterity in Parkinson's disease patients. *Informatics in Medicine Unlocked* **9,** 11–17 (2017).
17. Wissel, B. D. *et al.* Tablet-Based Application for Objective Measurement of Motor Fluctuations in Parkinson Disease. *Digit. Biomark.* **1,** 126–135 (2018).
18. Bot, B. M. *et al.* The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci. Data* **3,** 160011 (2016).
19. Olson, M., Lockhart, T. E. & Lieberman, A. Motor learning deficits in parkinson's disease (PD) and their effect on training response in gait and balance: A narrative review. *Front. Neurol.* **10,** 62 (2019).
20. Steib, S. *et al.* A single bout of aerobic exercise improves motor skill consolidation in parkinson's disease. *Front. Aging Neurosci.* **10,** 328 (2018).
21. Bryant, M. Effect of Resistance Exercise on Tremor and Hand Dexterity of Parkinson's Disease.
22. Krebs, H. I., Hogan, N., Hening, W., Adamovich, S. V. & Poizner, H. Procedural motor learning in Parkinson's disease. *Exp. Brain Res.* **141,** 425–437 (2001).
23. Sehm, B. *et al.* Structural brain plasticity in Parkinson's disease induced by balance training. *Neurobiol. Aging* **35,** 232–239 (2014).
24. Zhan, A. *et al.* High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection. *arXiv* (2016).
25. Mirelman, A. *et al.* Fall risk and gait in Parkinson's disease: the role of the LRRK2 G2019S mutation. *Mov. Disord.* **28,** 1683–1690 (2013).
26. Blin, O., Ferrandez, A. M. & Serratrice, G. Quantitative analysis of gait in Parkinson patients: increased variability of stride length. *J. Neurol. Sci.* **98,** 91–97 (1990).
27. Hausdorff, J. M. Gait dynamics in Parkinson's disease: common and distinct behavior among stride length, gait variability, and fractal-like scaling. *Chaos* **19,** 026113 (2009).
28. Miller Koop, M., Ozinga, S. J., Rosenfeldt, A. B. & Alberts, J. L. Quantifying turning behavior and gait in Parkinson's disease using mobile technology. *IBRO Rep.* **5,** 10–16 (2018).
29. Martinez-Mendez, R., Sekine, M. & Tamura, T. Postural sway parameters using a triaxial accelerometer: comparing elderly and young healthy adults. *Comput Methods Biomech Biomed Engin* **15,** 899–910 (2012).
30. Prieto, T. E., Myklebust, J. B., Hoffmann, R. G., Lovett, E. G. & Myklebust, B. M. Measures of postural steadiness: differences between healthy young and elderly adults. *IEEE Trans. Biomed. Eng.* **43,** 956–966 (1996).
31. Palakurthi, B. & Burugupally, S. P. Postural instability in parkinson's disease: A review. *Brain Sci.* **9,** (2019).
32. Pittman, B., Ghomi, R. H. & Si, D. Parkinson's Disease Classification of mPower Walking Activity Participants. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2018,** 4253–4256 (2018).
33. THEORY, B. & ALGORITHM. Implementing Positioning Algorithms Using Accelerometers.
34. Palmerini, L., Rocchi, L., Mellone, S., Valzania, F. & Chiari, L. Feature selection for accelerometer-based posture analysis in Parkinson's disease. *IEEE Trans. Inf. Technol.*

Biomed. **15,** 481–490 (2011).
35. Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A. E. & Moroz, I. M. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomed Eng Online* **6,** 23 (2007).
36. Chiaramonte, R. & Bonfiglio, M. Acoustic analysis of voice in Parkinson's disease: a systematic review of voice disability and meta-analysis of studies. *Rev. Neurol.* **70,** 393–405 (2020).
37. Rao, G. *et al.* Does this patient have Parkinson disease? *JAMA* **289,** 347–353 (2003).
38. Jobbágy, A., Harcos, P., Karoly, R. & Fazekas, G. Analysis of finger-tapping movement. *J. Neurosci. Methods* **141,** 29–39 (2005).
39. Cicchetti, D. V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* **6,** 284–290 (1994).

Table 1 Characteristics of study cohorts after data cleaning.

Task	HC		PD	
	Gender (M/F)	**Age (mean ± S.D)**	**Gender (M/F)**	**Age (mean ± S.D)**
Gait	655/152	49 ± 10.6	399/211	60.3 ± 8.9
Balance	668/155	48.9 ± 10.7	401/211	60.3 ± 8.9
Voice	1042/249	47.7 ± 10.4	571/322	60.1 ± 9
Tapping	1370/304	46.9 ± 10.1	630/340	59.9 ± 9

HC- Healthy Controls, PD- Parkinson's Disease, M- Male, F- Female, S.D- Standard Deviation

Figure 1 Effect size (Cohen's d). Best-performing features selected from different time points and repetitions. **(a)** Gait Task: FreezeInd- Freeze Index, PeakEnerg- Peak of Energy, skew- Skewness, MSI- Mean Stride Interval, RatioPower - Sum of the Power in the Freezing and Locomotor Band, frec_peak- Frequency at the Peak of Energy, iqr- Interquartile Range, cov- Coefficient of Variation, zcr- Zero-Crossing Rate, kur-Kurtosis, LB- Locomotor Band, FB- Freezing Band, a- accelerometer average signal, x- accelerometer Mediolateral signal, y- accelerometer vertical signal, z- accelerometer anteroposterior signal, **(b)** Balance Task: PeakEnergy - Peak of energy, TotalPower- Energy between 15-3.5 Hz, Power- Energy between 3.5-15Hz, rms- Root Mean Square, F50- Frequency Containing 50% of Total Power, FRQD- Frequency of Dispersion of the Power Spectrum, iqr- Interquartile Range, min- Minimum Value, CFREQ- Centroidal Frequency, RHL- Ratio Between Power in High Frequency and Low Frequency, MF- Medium Frequency (4-7Hz), VHF- Very High Frequency (>7Hz) , HF- Hight Frequency (>4Hz), LF- Low Frequency (0.15-3.5Hz), trem- tremor, post- postural, a- Accelerometer Average Signal, x- Accelerometer Mediolateral Signal, y- Accelerometer Vertical Signal, z- Accelerometer Anteroposterior Signal, Hz- hertz, **(c)** Voice Task: c_mean_MFCC1–12- Mean Value of Mel Frequency Cepstral Coefficients 1-12, gqc- Glottis Quotient Close, p95- 95th Percentile, shbd- Shimmer, **(d)** Tapping Task: iqr- Interquartile Range, TapInter- Tap Interval, buttonNoneFreq: Frequency of tapping outside the button, numberTaps- Number of Taps, DriftRight- Right Drift, corXY- Correlation of X and Y Positions, DriftLeft- Left Drift, mad- Median, min- Minimum, cv- Coefficient, Sd- standard deviation.

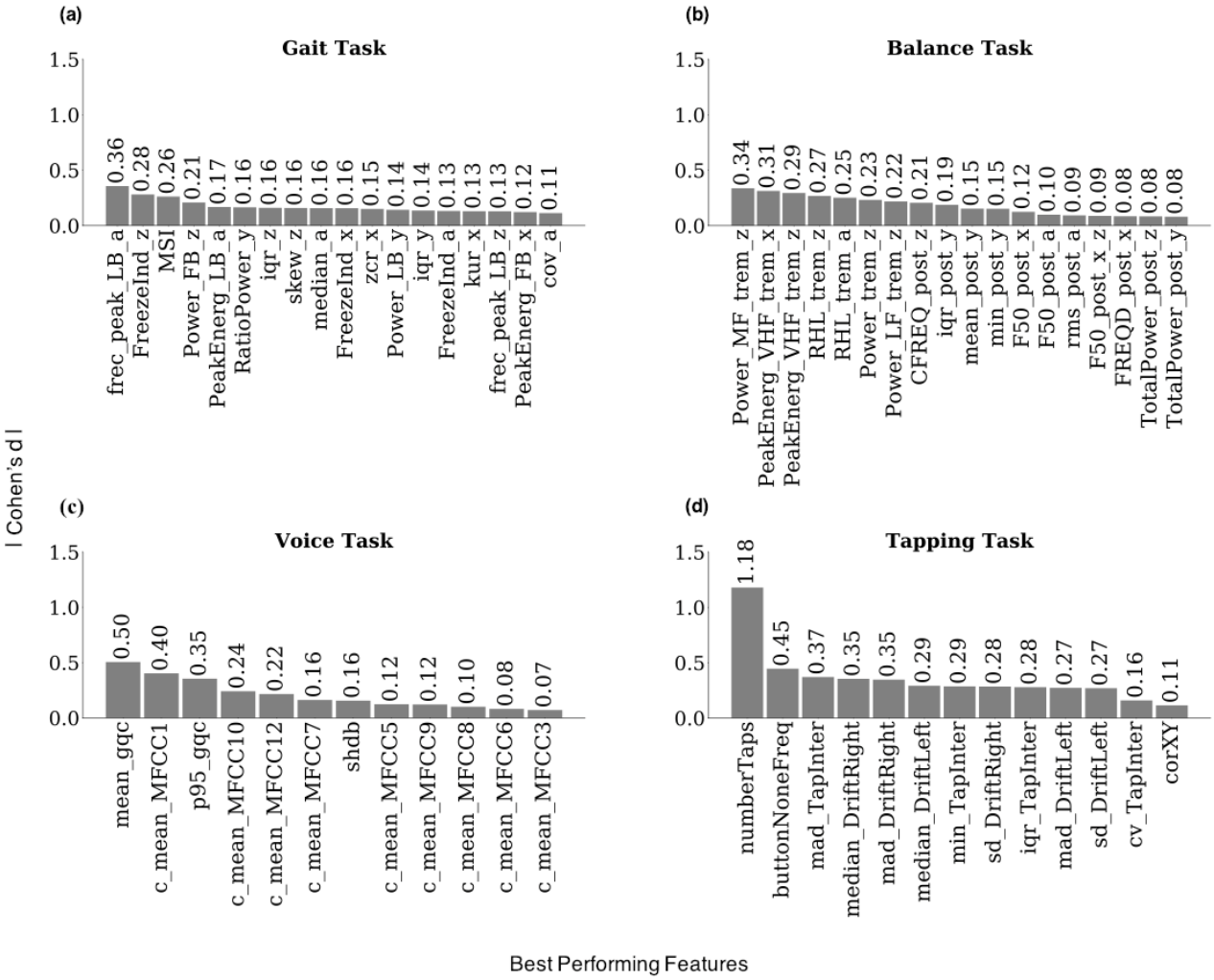
Figure 2 Median ICCs for the best performing features. **(a)** Median ICCs across different time points for the best performing features. **(b)** Median ICCs across different repetitions for the best performing features. Gait Task: FreezeInd- Freeze Index, PeakEnerg- Peak of Energy, skew- Skewness, MSI- Mean Stride Interval, RatioPower - Sum of the Power in the Freezing and Locomotor Band, frec_peak- Frequency at the Peak of Energy, iqr- Interquartile Range, cov- Coefficient of Variation, zcr- Zero-Crossing Rate, kur-Kurtosis, LB- Locomotor Band, FB- Freezing Band, a- accelerometer average signal, x- accelerometer Mediolateral signal, y- accelerometer vertical signal, z- accelerometer anteroposterior signal, Balance Task: PeakEnergy - Peak of energy, TotalPower- Energy between 15-3.5 Hz, Power- Energy between 3.5-15Hz, rms- Root Mean Square, F50- Frequency Containing 50% of Total Power, FRQD- Frequency of Dispersion of the Power Spectrum, iqr- Interquartile Range, min- Minimum Value, CFREQ- Centroidal Frequency, RHL- Ratio Between Power in High Frequency and Low Frequency, MF- Medium Frequency (4-7Hz), VHF- Very High Frequency (>7Hz) , HF- Hight Frequency (>4Hz), LF- Low Frequency (0.15-3.5Hz), trem- tremor, post- postural, a- Accelerometer Average Signal, x- Accelerometer Mediolateral Signal, y-

Accelerometer Vertical Signal, z- Accelerometer Anteroposterior Signal, Hz- hertz, Voice Task: c_mean_MFCC1–12- Mean Value of Mel Frequency Cepstral Coefficients 1-12, gqc- Glottis Quotient Close, p95- 95th Percentile, shbd- Shimmer, Tapping Task: iqr- Interquartile Range, TapInter- Tap Interval, buttonNoneFreq: Frequency of tapping outside the button, numberTaps- Number of Taps, DriftRight- Right Drift, corXY- Correlation of X and Y Positions, DriftLeft- Left Drift, mad- Median, min- Minimum, cv- Coefficient, Sd- standard deviation.

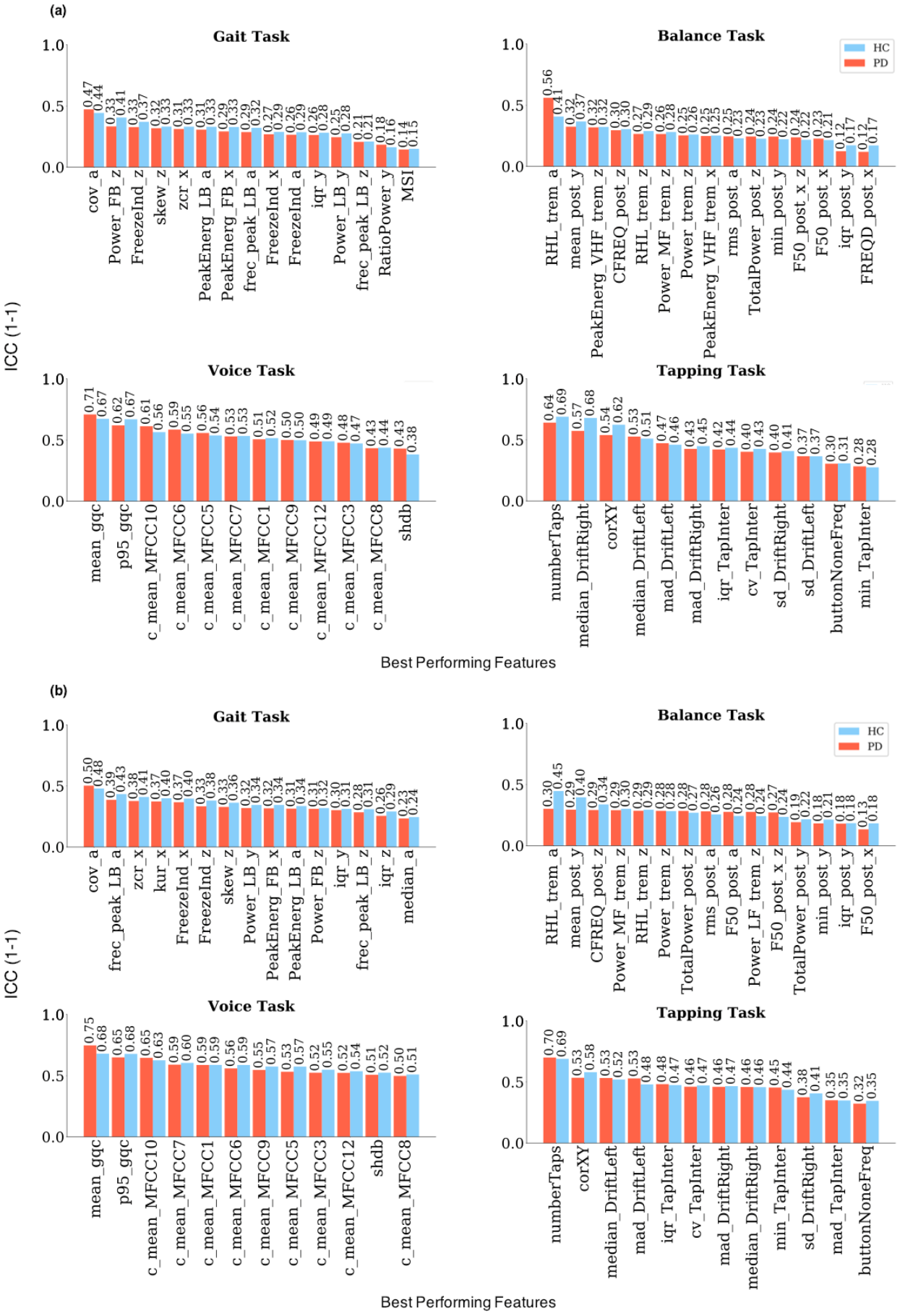
Figure 3 Mean value of the best performing features at baseline and across different time points, calculated for PD and HC separately. **(a)** Gait Task: FreezeInd- Freeze Index, PeakEnerg- Peak of Energy, skew- Skewness, MSI- Mean Stride Interval, RatioPower - Sum of the Power in the Freezing and Locomotor Band, frec_peak- Frequency at the Peak of Energy, iqr- Interquartile Range, cov- Coefficient of Variation, zcr- Zero-Crossing Rate, LB- Locomotor Band, FB- Freezing Band, a- accelerometer average signal,x- accelerometer Mediolateral signal, y- accelerometer vertical signal, z- accelerometer anteroposterior signal, **(b)** Balance Task: PeakEnergy - Peak of energy, TotalPower- Energy between .15-3.5 Hz, Power- Energy between 3.5-15Hz, rms- Root Mean Square, F50- Frequency Containing 50% of Total Power, FRQD- Frequency of Dispersion of the Power Spectrum, iqr- Interquartile Range, min- Minimum Value, CFREQ- Centroidal Frequency, RHL- Ratio Between Power in High Frequency and Low Frequency, MF- Medium Frequency (4-7Hz), VHF- Very High Frequency (>7Hz) , HF- Hight Frequency (>4Hz), LF- Low Frequency (0.15-3.5Hz), trem- tremor, post- postural, a- Accelerometer Average Signal, x- Accelerometer Mediolateral Signal, y- Accelerometer Vertical Signal, z- Accelerometer Anteroposterior Signal, Hz- hertz, **(c)** Voice task : c_mean_MFCC1–12- Mean Value of Mel Frequency Cepstral Coefficients 1-12, gqc- Glottis Quotient Close, p95- 95th Percentile, shbd- Shimmer, **(d)** Tapping Task: iqr- Interquartile Range, TapInter- Tap Interval, buttonNoneFreq: Frequency of tapping outside the button, numberTaps- Number of Taps, DriftRight- Right Drift, corXY- Correlation of X and Y Positions, DriftLeft- Left Drift, mad- Median, min- Minimum, cv- Coefficient, Sd- standard deviation.

Figure 4 Mean value of the best performing features at baseline and across different time points, calculated for PD and HC separately. **(a)** Gait Task: FreezeInd- Freeze Index, PeakEnerg- Peak of Energy, frec_peak- Frequency at the Peak of Energy, skew- Skewness, iqr- Interquartile Range, cov- Coefficient of Variation, zcr- Zero-Crossing Rate, kur-Kurtosis, LB- Locomotor Band, FB- Freezing Band, a- accelerometer average signal,x- accelerometer Mediolateral signal, y- accelerometer vertical signal, z- accelerometer anteroposterior signal, **(b)** Balance Task: PeakEnergy - Peak of energy, TotalPower- Energy between .15-3.5 Hz, Power- Energy between 3.5-15Hz, rms- Root Mean Square, F50- Frequency Containing 50% of Total Power, FRQD- Frequency of Dispersion of the Power Spectrum, iqr- Interquartile Range, min- Minimum Value, CFREQ- Centroidal Frequency, RHL- Ratio Between Power in High Frequency and Low Frequency, MF- Medium Frequency (4-7Hz), VHF- Very High Frequency (>7Hz) , HF- Hight Frequency (>4Hz),

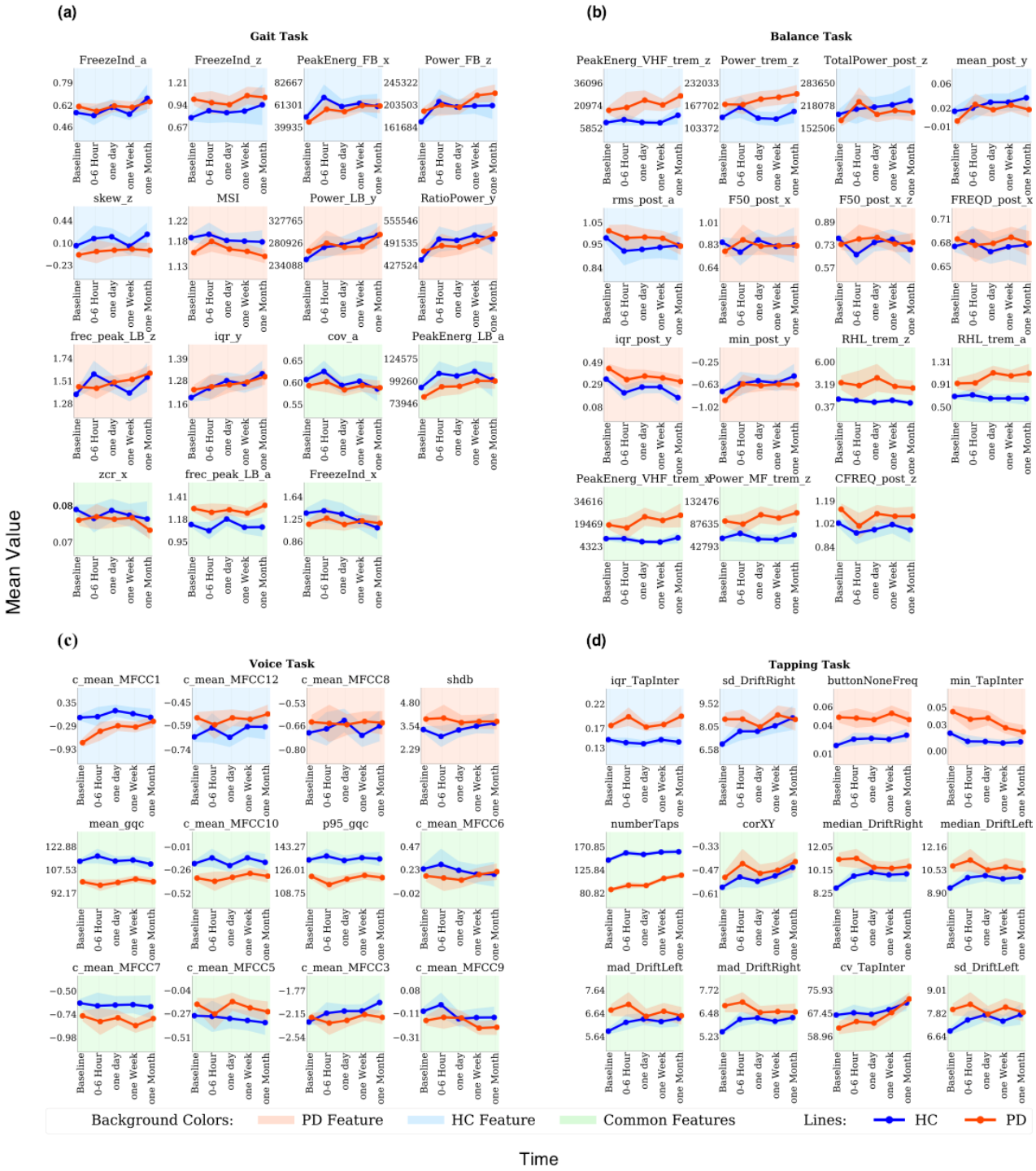
LF- Low Frequency (0.15-3.5Hz), trem- tremor, post- postural, a- Accelerometer Average Signal, x- Accelerometer Mediolateral Signal, y- Accelerometer Vertical Signal, z- Accelerometer Anteroposterior Signal, Hz- hertz, **(c)** Voice Task: c_mean_MFCC1–12- Mean Value of Mel Frequency Cepstral Coefficients 1-12, shbd- Shimmer, gqc- Glottis Quotient Close, p95- 95th Percentile, **(d)** Tapping Task: Tapping Task: iqr- Interquartile Range, TapInter- Tap Interval, buttonNoneFreq: Frequency of tapping outside the button, numberTaps- Number of Taps, DriftRight- Right Drift, corXY- Correlation of X and Y Positions, DriftLeft- Left Drift, mad- Median, min- Minimum, cv- Coefficient, Sd- standard deviation.


(a)
Gait Task
1.5
1.0
0.5
0.0
0.36
0.28
0.26
0.21
0.17
0.16
0.16
0.16
0.16
0.16
0.15
0.14
0.14
0.13
0.13
0.13
0.12
0.11
frec_peak_LB_a
FreezeInd_z
MSI
Power_FB_z
PeakEnerg_LB_a
RatioPower_y
iqr_z
skew_z
median_a
FreezeInd_x
zcr_x
Power_LB_y
iqr_y
FreezeInd_a
kur_x
frec_peak_LB_z
PeakEnerg_FB_x
cov_a
(b)
Balance Task
1.5
1.0
0.5
0.0
0.34
0.31
0.29
0.27
0.25
0.23
0.22
0.21
0.19
0.15
0.15
0.12
0.10
0.09
0.09
0.08
0.08
0.08
Power_MF_trem_z
PeakEnerg_VHF_trem_x
PeakEnerg_VHF_trem_z
RHL_trem_z
RHL_trem_a
Power_trem_z
Power_LF_trem_z
CFREQ_post_z
iqr_post_y
mean_post_y
min_post_y
F50_post_x
F50_post_a
rms_post_a
F50_post_x_z
FREQD_post_x
TotalPower_post_z
TotalPower_post_y
(c)
Voice Task
1.5
1.0
0.5
0.0
0.50
0.40
0.35
0.24
0.22
0.16
0.16
0.12
0.12
0.10
0.08
0.07
mean_gqc
c_mean_MFCC1
p95_gqc
c_mean_MFCC10
c_mean_MFCC12
c_mean_MFCC7
shdb
c_mean_MFCC5
c_mean_MFCC9
c_mean_MFCC8
c_mean_MFCC6
c_mean_MFCC3
(d)
Tapping Task
1.5
1.0
0.5
0.0
1.18
0.45
0.37
0.35
0.35
0.29
0.29
0.28
0.28
0.27
0.27
0.16
0.11
numberTaps
buttonNoneFreq
mad_TapInter
median_DriftRight
mad_DriftRight
median_DriftLeft
min_TapInter
sd_DriftRight
iqr_TapInter
mad_DriftLeft
sd_DriftLeft
cv_TapInter
corXY
| Cohen's d |
Best Performing Features

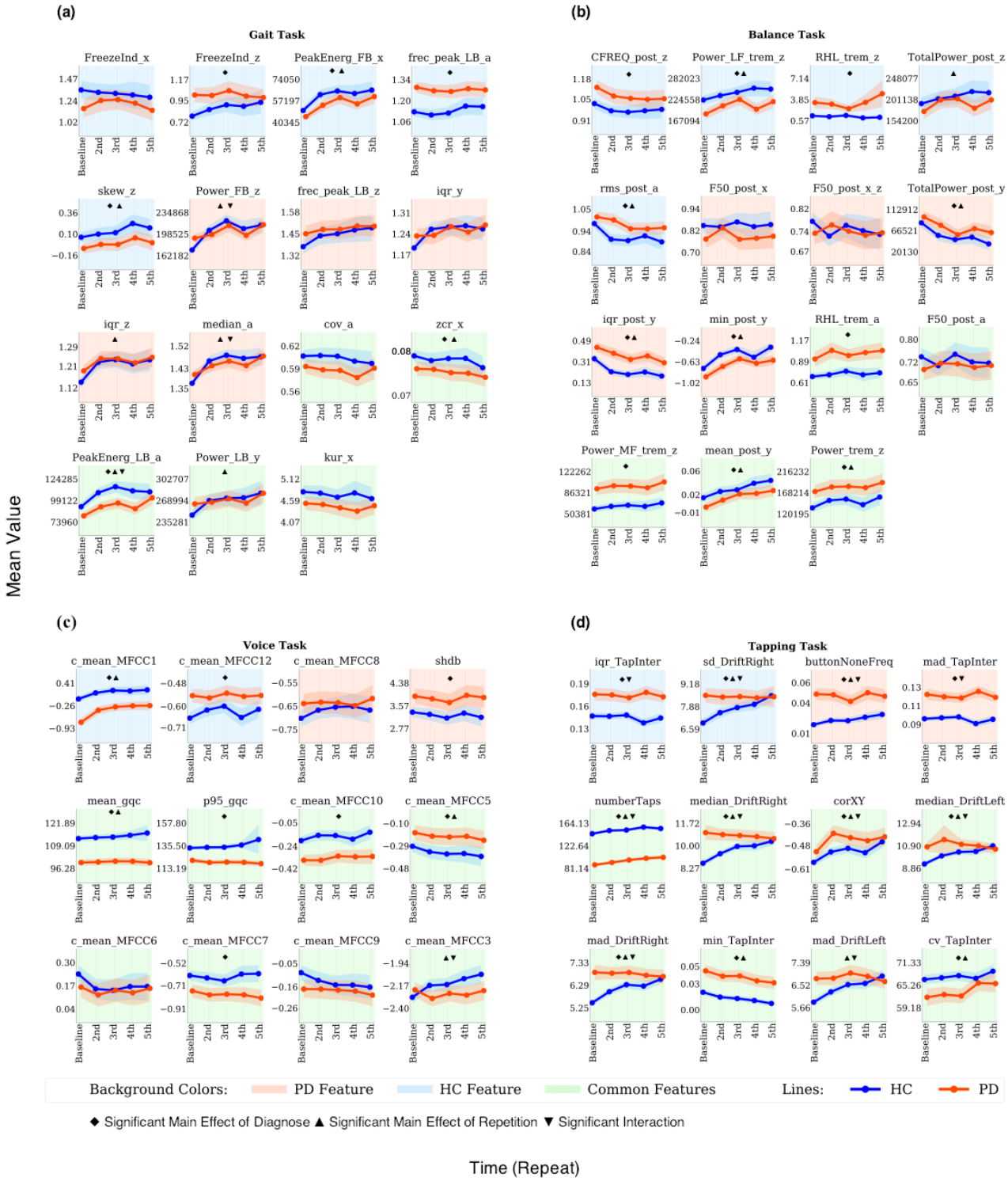
[Figure 1]


(a)
Gait Task
Balance Task
Voice Task
Tapping Task
HC
PD
ICC (1-1)
Best Performing Features
(b)
Gait Task
Balance Task
Voice Task
Tapping Task
HC
PD
ICC (1-1)
Best Performing Features

[Figure 2]


(a)
Gait Task
FreezeInd_a
FreezeInd_z
PeakEnerg_FB_x
Power_FB_z
skew_z
MSI
Power_LB_y
RatioPower_y
frec_peak_LB_z
iqr_y
cov_a
PeakEnerg_LB_a
zcr_x
frec_peak_LB_a
FreezeInd_x
(b)
Balance Task
PeakEnerg_VHF_trem_z
Power_trem_z
TotalPower_post_z
mean_post_y
rms_post_a
F50_post_x
F50_post_x_z
FREQD_post_x
iqr_post_y
min_post_y
RHL_trem_z
RHL_trem_a
PeakEnerg_VHF_trem_x
Power_MF_trem_z
CFREQ_post_z
(c)
Voice Task
c_mean_MFCC1
c_mean_MFCC12
c_mean_MFCC8
shdb
mean_gqc
c_mean_MFCC10
p95_gqc
c_mean_MFCC6
c_mean_MFCC7
c_mean_MFCC5
c_mean_MFCC3
c_mean_MFCC9
(d)
Tapping Task
iqr_TapInter
sd_DriftRight
buttonNoneFreq
min_TapInter
numberTaps
corXY
median_DriftRight
median_DriftLeft
mad_DriftLeft
mad_DriftRight
cv_TapInter
sd_DriftLeft
Baseline
0-6 Hour
one day
one Week
one Month
Mean Value
Time
Background Colors:
PD Feature
HC Feature
Common Features
Lines:
HC
PD

[Figure 3]


(a)
Gait Task
FreezeInd_x
FreezeInd_z
PeakEnerg_FB_x
frec_peak_LB_a
skew_z
Power_FB_z
frec_peak_LB_z
iqr_y
iqr_z
median_a
cov_a
zcr_x
PeakEnerg_LB_a
Power_LB_y
kur_x
(b)
Balance Task
CFREQ_post_z
Power_LF_trem_z
RHL_trem_z
TotalPower_post_z
rms_post_a
F50_post_x
F50_post_x_z
TotalPower_post_y
iqr_post_y
min_post_y
RHL_trem_a
F50_post_a
Power_MF_trem_z
mean_post_y
Power_trem_z
(c)
Voice Task
c_mean_MFCC1
c_mean_MFCC12
c_mean_MFCC8
shdb
mean_gqc
p95_gqc
c_mean_MFCC10
c_mean_MFCC5
c_mean_MFCC6
c_mean_MFCC7
c_mean_MFCC9
c_mean_MFCC3
(d)
Tapping Task
iqr_TapInter
sd_DriftRight
buttonNoneFreq
mad_TapInter
numberTaps
median_DriftRight
corXY
median_DriftLeft
mad_DriftRight
min_TapInter
mad_DriftLeft
cv_TapInter
Baseline
2nd
3rd
4th
5th
Mean Value
Background Colors: PD Feature HC Feature Common Features
Lines: HC PD
◆ Significant Main Effect of Diagnose ▲ Significant Main Effect of Repetition ▼ Significant Interaction
Time (Repeat)

[Figure 4]