from the Institute for Neuroscience and Medicine
at the Heinrich Heine University Düsseldorf

Sensor-based Assessments as a Disease Progression and Treatment Biomarker for Neuropsychiatric Diseases

Dissertation

to obtain the academic title of Doctor of Philosophy (PhD) in Medical Sciences

from the Faculty of Medicine at Heinrich Heine University Düsseldorf

submitted by

Mehran Sahandi Far

(2023)

„Als Inauguraldissertation gedruckt mit Genehmigung der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

Gez.:

Dekan: Prof. Dr. med. Nikolaj Klöcker

Gutachter/innen: Prof. Dr. Juergen Dukart, Prof. Dr. Oliver Kuß

Parts of this work have been accepted and published:

1) Sahandi Far, Mehran, Michael Stolz, Jona M. Fischer, Simon B. Eickhoff, and Juergen Dukart. "JTrack: a digital biomarker platform for remote monitoring of daily-life behaviour in health and disease." Frontiers in Public Health 9 (2021): 763621.

2) Rentz, Clara, Mehran Sahandi Far, Maik Boltes, Alfons Schnitzler, Katrin Amunts, Juergen Dukart, and Martina Minnerop. "System Comparison for Gait and Balance Monitoring Used for the Evaluation of a Home-Based Training." Sensors 22, no. 13 (2022): 4975.

3) Sahandi Far, Mehran, Simon B. Eickhoff, Maria Goni, and Juergen Dukart. "Exploring test-retest reliability and longitudinal stability of digital biomarkers for Parkinson disease in the m-power data set: Cohort study." Journal of Medical Internet Research 23, no. 9 (2021): e26608.

4) Goñi, María, Simon B. Eickhoff, Mehran Sahandi Far, Kaustubh R. Patil, and Juergen Dukart. "Smartphone-based digital biomarkers for Parkinson’s disease in a remotely-administered setting." IEEE access 10 (2022): 28361-28384.

Summary (German)

Der wachsende Trend zur individuellen Gesundheitsversorgung und Digitalisierung im Gesundheitswesen, z.B. mit Hilfe von *remote monitoring* (Fernüberwachung), hat zu einem verstärkten Interesse an der Verwendung integrierter Sensoren in bspw. Smartphones und Smartwatches, in klinischen Studien geführt. Die von solchen Geräten erfassten gesundheitsbezogenen Daten werden als digitale Biomarker (DB) bezeichnet. Im Gegensatz zu bisherigen klinischen Evaluationsmethoden liefern DB kostengünstige, objektive und ökologisch valide Daten. Mit Hilfe von DB ist es besser möglich, innerhalb klinischer Studien, eine größere und vielfältigere Population zu untersuchen. Darüber hinaus liefern DB Daten mit hoher zeitlicher und räumlicher Auflösung, was zu einem besseren Verständnis des Krankheitsverlaufs und -fortschritts beitragen kann.

Aufgrund des Mangels an objektiven Bewertungsinstrumenten, stehen neurodegenerative Krankheiten im Fokus der DB-Forschung. Für die Erforschung DB eignet sich die Parkinson Krankheit (PD), hinsichtlich des heterogenen Eintrittsalters, der Symptomprävalenz, der Schweregradprogression und der vielseitigen Symptome besonders gut. Es erscheint notwendig, die Verwendung DB in die klinische Diagnostik von Parkinson zu integrieren. Diese Integration, v.a. in der häuslichen Umgebung, stellt jedoch eine praktische Herausforderungen dar. Darüber hinaus ist die longitudinale Stabilität von DB, die innerhalb solcher Umgebungen erhoben wurden, noch nicht ausreichend erforscht worden - bisherige Studien beschränken sich auf Laborsituationen. Ziel dieser Arbeit ist es, Erkenntnisse darüber zu erlangen, wie *remote monitoring* in einer häuslichen Umgebung in Kombination mit einer Plattform zur Datenerhebung genutzt werden kann, um die Evaluation von Parkinson zu verbessern.

Der erste Teil der Arbeit befasst sich mit der Plattform „JTrack", die für das *remote monitoring* von Krankheiten entwickelt wurde. „JTrack“ eignet sich dafür, mehrere Aspekte eines Krankheitsbildes zu erfassen und stellt somit ein umfassendes Messinstrument für klinische Studien dar. Zudem wird die Übereinstimmung der mit „JTrack" erfassten Daten mit zwei gängigen stationären Systemen zur Analyse von Gang und Gleichgewicht bewertet, um das Potenzial des Einsatzes von Smartphones und insbesondere von „JTrack" in künftigen klinischen Studien darzulegen.

Im zweiten Teil wird beschrieben, inwieweit häufig berichtete Merkmale von PD, als Biomarker verwendet werden können. Zu diesem Zweck wurde zunächst die Test-Retest-Zuverlässigkeit und die longitudinale Stabilität dieser Merkmale untersucht. Anschließend wurden Algorithmen des maschinellen Lernens verwendet, um zu bewerten, inwiefern diese Merkmale verwendet werden können, um zwischen PD und gesunden Kontrollprobanden zu unterscheiden. Außerdem wurde der Einfluss verschiedener Störfaktoren wie Komorbiditäten, Alter und Geschlecht auf die Vorhersageleistung der maschinellen Lernalgorithmen untersucht. Dazu wurden verschiedenen Aufgaben (z.B. Gang, Gleichgewicht) der m-Power-Datenbank verwendet, die im unkontrollierten, häuslichen Umfeld von den Patienten durchgeführt und mit Hilfe des *remote monitorings* erfasst wurden.

Insgesamt werden innerhalb dieser Arbeit die Möglichkeiten und Grenzen der Verwendung von Smartphones für die Diagnose von PD diskutiert. Es werden zudem mögliche beeinflussende Faktoren im Zusammenhang mit DB bei Fern- und selbstverwalteten Erfassungsmethoden dargestellt. Es wird außerdem verdeutlicht, dass kontrolliertere, standardisiertere, empfindlichere und zuverlässigere DB entwickelt werden müssen, bevor sie in klinischen Anwendungen (Apps) eingesetzt werden sollten. Schließlich wird eine neue DB-Plattform für das *remote monitoring* vorgestellt, die für verschiedene Arten von Krankheiten genutzt werden kann.

Summary

The growing trend of personalised health care and remote monitoring has led to increased interest in using embedded sensors in portable smart devices (smartphones and smartwatches) in clinical studies. Health-related data collected from such devices are referred to as Digital Biomarkers (DBs). Unlike traditional in-clinic assessment methods, DBs provide cost-effective, objective, and ecologically valid data. DBs enable clinical studies to recruit a larger and more diverse population. Furthermore, DBs provide high temporal and spatial resolution data, which increase the chance of gaining a comprehensive understanding of disease progression.

Neurodegenerative diseases, due to their lack of accessible and objective assessment tools, have been a primary focus for the DBs research community. Parkinson's disease (PD) is particularly well-suited for studying DBs due to its heterogeneous onset age, symptom prevalence, severity progression rate, and multiple aspects of the disease. Therefore, there is a need to integrate DBs and remote assessment into the routine clinical evaluation of PD. However, using DBs for PD in non-controlled, at-home settings poses practical challenges that have hindered this goal. Additionally, the longitudinal stability of DBs collected in such settings has not yet been thoroughly investigated, with previous studies limited to in-lab settings. Thus, this thesis aims to provide insight into how remote monitoring in an at-home environment alongside the data collection methods can be leveraged to improve the way PD is assessed.

The first section of this dissertation focuses on introducing a platform named "JTrack", designed for remote disease monitoring and to address technical aspects such as security, privacy, modularity, and reusability. This platform aims to provide a comprehensive solution for clinical studies involving multiple aspects of various diseases. In addition, this section assesses the agreement between features collected through "JTrack" with two widely used stationary systems for analysing gait and balance, demonstrating the potential of using smartphones and particularly the "JTrack" platform in future clinical studies.

The second part of this thesis investigates the potential of using various commonly reported features in PD studies as biomarkers. To do this, we first investigate these features' test-retest reliability and longitudinal stability, considering how the timescale may affect their stability. Next, we use various machine learning algorithms to assess the ability of these features to differentiate between PD and HC. Also, we evaluated the influence of different confounding factors such as comorbidities, age, and sex on the prediction performance of the machine learning algorithms. For this, the various tasks (gait, balance, voice, and tapping) of the m-Power database, collected remotely and in a self-managed setting, were investigated.

Overall, this thesis discusses the potential and limitations of using smartphones for remote assessment of PD. It examines the possible sources of confounding factors related to DBs in remote and self-managed collection methods. It also highlights the need to develop more controlled, standardised, sensitive, and reliable DBs before taking them into any clinical application. This thesis also introduces a new DBs platform for remote assessment, which can be leveraged for various types of disease.

List of Abbreviations

DB	Digital Biomarker
PD	Parkinson's Disease
HC	Healthy Control
GDPR	General Data Protection Regulation
MDS-UPDRS	Movement Disorder Society's Unified Parkinson's Disease Rating Scale
IMU	Inertial Measurement Unit

Table of Contents

Introduction

The challenges (e.g., time constrain, accessibility, cost, limited availability, and infrequent visits) associated with current in-clinic visit assessment methods have motivated the development of novel disease assessment techniques using sensor-embedded smart devices. Various sensors have recently been incorporated into wearables and smartphones. This rich combination of built-in sensors, processing power, wireless connection capability, and customisable applications opens the door to collecting objective and quantifiable health-related data. The data acquired via smart devices in clinical studies are referred to as digital biomarkers (DBs) (1–3). DBs can provide cost-effective, objective, and ecologically valid health-related data. Furthermore, real-time longitudinal data collection from a larger population allows a better understanding of disease evolution and variation between individuals.

In clinical trials, two major deployment models for DBs can be outlined. The first, consists of samples obtained in a controlled environment, for example, a clinical or movement laboratory, using devices designed for this particular purpose (4,5). The second approach brings together commercially accessible hardware (such as smartphones and smartwatches) with custom applications to be used in loosely controlled and self-administrated in-home settings (6–8). The main benefit of the second approach is the capacity to monitor patients in (near) real-time remotely and for extended periods of time, such as months and years. Additionally, smartphones are practical tools for crowdsourcing due to their widespread use. Although there are already many platforms designed for at-home disease monitoring (9–13), the majority of them have several shortcomings, including a lack of memory and energy optimisation, an easy-to-use user interface, security, and privacy, and a lack of availability in official application distribution stores. One of the main challenges posed by these platforms is maintaining the confidentiality of the collected data (14), which is consistently targeted in cyberattacks. Therefore, it is strongly recommended that no identifying information be collected (or use proper anonymising and encryption methods) in the first place. Another challenge is the large amount of data generated, which requires the creation of efficient data storage and streaming methods, not only at the device level but also at the storage endpoint. Smartphones often have limited internal storage capacity and access to low-cost (free) connectivity, so it is important to have automatic synchronisation strategies to maximise memory usage efficiency.

Additionally, the flexible and modular design of DBs platforms to accommodate a wide variety of sensor types is another key feature that should not be overlooked. These platforms should allow researchers and clinicians to easily add or remove different types of sensors (data sources). However, most of the existing platforms are designed to focus on a specific aspect of a disease, which negatively affects their reusability and flexibility. Nevertheless, there is a gap between the tools available and needed for this purpose, which motivated us to introduce our platform, "JTrack".

DBs have been used in the assessment of various diseases, from heart disease and diabetes to cancers and chronic pain. In particular, DBs have been extensively used to measure and monitor the progression of certain neurodegenerative disorders, such as Parkinson's Disease (PD), one of the very well-suited neurodegenerative diseases for DBs studies (6). Given different aspects of the disease such as motor, speech, and sensory disturbances which are highly heterogeneous across patients and disease stages (15), remote monitoring in PD is a powerful way to better understand the feasibility of deploying such methods as a complement to traditional in-clinic visits (6,16).

PD is the second most prevalent neurodegenerative disorder of the elderly, involving approximately 9.4 million people over the age of 60 worldwide in 2020 (17). This complex progressive movement disease is essentially characterised by motor signs, such as muscle rigidity, rest tremor, akinesia, and postural instability (18). However, diverse non-motor symptoms, including sleep disorders, psychiatric disorders, and sensory disturbances are also present in PD (18,19).

PD is classically characterised by a gradual loss of dopaminergic neurons in the substantia nigra, a region of the midbrain responsible for movement control, as well as an abnormal accumulation of Alpha-synuclein (α-synuclein) aggregates called Lewy bodies, which blocks dopamine's production and transmission, and subsequent loss of dopamine in the striatum (20–22). Nevertheless, increasing evidence has apparently shown that PD is a multisystem disorder and the substantia nigra is not the only or the first brain region damaged in PD. The degenerative process in PD is much more extensive that affects the whole nervous system (21,23).

From a neuropathological point of view, while the motor symptoms of PD are mainly attributed to the degeneration of dopaminergic neurons in the nigrostriatal system, the non-motor symptoms such as sleep abnormality, loss of smell, depression, and mood disorders are associated with a start of α-synuclein aggregates beyond the nigrostriatal dopaminergic system (21,24). Furthermore, the symptomatic phase of PD, which is mainly characterised by the above-mentioned motor symptoms, is not evident until approximately 60–70% of dopaminergic neurons in the substantia nigra have already degenerated (25), whereas the prodromal stage of the disease is often characterised by non-motor symptoms which may precede the appearance of characteristic motor symptoms by several years (21).

The exact aetiology of PD is unknown to date, but several genetic risk factors have now been characterised, which cause a rare familial form of PD (26). There is also some evidence that environmental factors (i.e., viruses and bacteria, toxic chemicals, heavy metals, and free radicals) may cause the deterioration of dopamine-producing neurons, leading to the development of PD, although the role of these remains unclear (27). Furthermore, recent studies suggest that viral or chemical exposure may cause inflammation in the olfactory system and gut leading to the initial α-synuclein misfolding, aggregation, and propagation to the brain (28).

Treatment of PD predominantly focuses on symptomatic relief with drugs aiming to either restore the level of dopamine in the striatum or to act on dopamine receptors (19,20). Although many other drugs are also being used to target specific symptoms, the non-motor symptoms of PD often go unrecognised and therapeutic management of these symptoms remains challenging, negatively impacting patients' quality of life (19).

Finally, PD is one of the world's fastest-growing neurological disorders, yet much is unknown about its current global economic burden. The direct and indirect burden of PD is estimated to exceed $79 billion in the US by 2037 (for an estimated 1.6 million patients) (29). The prevalence of PD increases with age, which is considered the most significant risk factor for the development and progression of PD (27). This means that society and the economy will be confronted with severe challenges as the proportion of older people in the total population continues to rise over the coming decade. Therefore, taking new approaches for early diagnosis and accurate interventions and monitoring the progression is needed to improve the efficiency and accuracy of treatments in patients suffering from PD.

Despite the recent interest and novel approaches to facilitate the diagnosis of PD, there are still insufficient biomarkers to assist this process. The frequently used methods often rely on in-clinic visits and subjective evaluation of patients combining clinician-rated systems such as the Movement Disorder Society's Unified Parkinson's Disease Rating

Scale (MDS-UPDRS) and self-reported information (30,31). These in-person visits are limited in terms of travel distance and distribution of specialised physicians and professional healthcare personnel which are often expensive and inaccessible to most people. Furthermore, the quality and quantity of observations that these approaches produce are frequently inadequate, and they also have a strong potential for a large degree of inter- and intra-rater variability (8,32,33). In addition, they are affected by several different recall biases, which can introduce errors and limit the usefulness of findings based on these reports (34). Nevertheless, these restrictions result in poor diagnostic accuracy for PD, particularly in the early stages of the disease, and this continues to be a significant impediment to patient care (35).

Digital sensors such as Inertial Measurement Unit (IMU) have been widely used for motion analysis and are the most frequently used sensor in PD studies (33). Due to the fact that a large number of motor-UPDRS questions are directly related to particular gait and postural instability symptoms, this domain has received the most attention from IMU sensor-based research. Unlike more complex methods such as walking carpets, IMU sensors provide high-resolution data with low cost and ease of use.

Most of the research on DBs for PD focuses on discovering features that can differentiate between PD and Healthy Control (HC) and correlating these features to clinical scores such as the MDS-UPDRS. In addition, it is desired to have clinically accepted and interpretable features that can also identify symptom severity and medication state (on/off) (for a detailed review see (33)).

In this regard, features such as the number and symmetry of steps, gait speed, cadence, gait variability, and freezing of gait are among the most frequently reported features for gait analysis (5,36–44). The jerkiness of posture, high-frequency power, and total distance moved are among the features extracted for postural instability assessment (38,45–48). Other features extracted from frequency and spatiotemporal-related aspects of a signal are also reported frequently.

IMU sensors are also frequently used to analyse tremor-related symptoms, and features such as temporal and frequency domain-related features, entropy, root mean square, and signal amplitude are widely reported (2,4,49–51). IMU sensors, combined with touchscreens, are also commonly used to assess bradykinesia-related symptoms and the most common features are the number of taps, speed of tap, and temporal variability of taps as frequency domain-oriented features (49,50,52). In excess of motor symptoms, voice dysfunction is a frequently reported symptom related to PD (53) that may be caused by the rigidity or bradykinesia of the laryngeal muscles in PD that can result in several abnormalities, including reduced voice volume, breathy voice, poor articulation, jerky speech, and incomplete glottic closure. (53–56). The embedded microphone in smartphones makes voice assessment simple nowadays and features such as Mel-Frequency-Cepstral-Coefficients, Harmonics-to-Noise-Ratio, shimmer, and jitter are among the features that different researchers have reported (15,54,57–61).

Several tasks are designed to aid the objective assessment of PD symptoms using smartphones, such as spirography, sway task, tapping task, and Time-Up-and-Go (TUG). These tasks can also be used with passive monitoring of daily activities, which does not require user interaction with an application. Passive data collection may provide comprehensive insights into PD symptoms through features such as daily activity, time-in-bed, and the number of sit-to-stand transitions (62,63). However, passive monitoring approaches also necessitate the investigation of advanced algorithms to remove a substantial amount of noise and classify huge amounts of unlabelled data (64).

The use of digital sensors in clinical settings has demonstrated their feasibility in objectively assessing PD symptoms (2). However, the majority of these studies were conducted in a controlled in-clinic environment with a small cohort size (6), which limits the diversity of the study cohort in terms of disease stages and severity of symptoms. Moreover, these studies often lack longitudinal data with higher temporal resolution. Age and disease duration are the two major timescales that affect PD's clinical manifestation (17), and longitudinal studies of PD, particularly in those who do not yet exhibit any motor symptoms but only pre-motor symptoms, would play a key role in gaining a more accurate insight of disease course and onset. To address these issues and move toward personalised health, it is necessary to shift toward using at-home and self-managed data collection methods in longitudinal studies with large and diverse populations.

However, many sources of variability can influence the validity and reliability of DBs, leading to invalid results and hindering the interpretation and translation of clinical trials. These sources of variability include sensor-level (measurement) variability (i.e., recording frequency, placement, quality, etc.) and individual-level variability (i.e., gender, medication, disease manifestation, etc.). The use of at-home-based data collection introduces an additional source of variabilities, such as selection bias, the effect of learning and motivation, the inclusion of multiple recordings for each participant, and the ability to follow instructions (6,65). Furthermore, there is still a lack of clarity on whether the impact of these variations remains the same throughout the course of the disease or whether they have varying effects. When taking into consideration the heterogeneous character of PD, this challenge also receives an additional layer of difficulty (64). In addition, the impact of these variabilities when using the machine learning approaches investigated in automated PD classification—which may result in overly optimistic results—has not been addressed properly thus far.

To address these open questions and examine the concepts introduced above, the m-Power dataset, assessed outside of clinical environments in self-administered settings, is being investigated (66).

Ethics protocols

The ethics protocols were approved by the Ethics Committee of Heinrich Heine University Düsseldorf (Studien-Nr.: 2020-1077-andere Forschung erstvotierend
Comparison of passive monitoring and self-reported social behaviour using smartphone based assessments), And the ethics committee of the Psychology faculty of the Heinrich Heine University Düsseldorf. (DU01-2021-01).

Aims and Organization of this Thesis

Due to the shortcomings and difficulties of current in-clinic visits, there is currently an active interest in leveraging DBs obtained from smartphones and wearable technology in the remote assessment of diseases like PD. These biomarkers can be used to classify diseases or assess the severity of their symptoms, and at-home assessment methods aim to maximise ecological validity and provide improved insight into disease progress and onset (67). Thus, in this thesis, I aim to contribute to this field by pursuing the following steps. First, I introduce a reusable and open-source platform to collect research-grade context-driven data while prioritising privacy and security concerns. Second, in a collaboration project, we assessed the agreement of data collected from this platform with advanced stationary platforms (force plate and motion capture systems), which are accepted as the gold standard for motion analysis. Third, I evaluated the longitudinal sensitivity and test-retest reliability of commonly reported features extracted from the m-Power dataset—the remote study of PD in a self-administrated at-home protocol and then assessed if the longitudinal behaviour of these potential DBs differed between disease groups and if this fluctuation was related to disease or other confounding factors that may

have had a contribution. Finally, we investigated these potential DBs to evaluate the sensitivity and specificity of different machine learning algorithms on the differentiation of PD from HC and what and how confounding factors affect this performance.
The following articles were published as part of my doctoral work and represent the main structure of this dissertation.

Study 1: In the first study, I addressed and aimed to fill the gap in the trade-off between privacy, optimisation, stability, and research-grade data quality that were not well met in previously introduced platforms where we're able to collect context-driven data. In this study, the "JTrack" platform was introduced as a single solution for digital phenotyping. "JTrack" comprises a smartphone application and an online dashboard enabling remote data collection and study management. It gives a flexible and modular environment for collecting sensor and smartphone usage data, with privacy and General Data Protection Regulation (GDPR) compliance as top priorities.

Study 2: In this study, we assessed the feasibility of using smartphones for gait and balance analysis. In this context, we examined the agreement between gait and balance features derived using smartphones and two widely used stationary gait analysis systems (e.g., force plate and motion capture systems) that are considered the gold standard in this field. In particular, we intend to assess the viability of adopting smartphone-based gait and balance studies in lieu of advanced techniques designed for laboratory use.

Study 3: The third study was devoted to investigating the longitudinal stability and test-retest reliability of various features extracted in Study 4. Despite its importance, little attention has been paid to evaluating the test-retest reliability and longitudinal stability of DBs in a loosely controlled self-administered setting. Therefore, in this study, I invested different sources of variation in the long-term performance of DBs such as the repetition, learning, and medication effects.

Study 4: In this study, we investigated various features derived from common tasks (i.e., gait, balance, voice, tapping) of the m-Power database performed in a self-administered setting for a remote assessment of PD. In this context, we have evaluated the specificity and sensitivity of various machine learning algorithms in differentiating PD and HC. We subsequently explored the influence of different confounding variables such as age, sex, and comorbidities on the classification performance for each task and the combination of all tasks.

Paper1: JTrack: A Digital Biomarker Platform for Remote Monitoring of Daily-Life Behavior in Health and Disease

Mehran Sahandi Far[1,2], Michael Stolz[1], Jona M. Fischer[1], Simon B Eickhoff[1,2], Juergen Dukart[1,2]

[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

Corresponding Author:

Juergen Dukart, PhD
Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7) Research Centre Jülich
Wilhelm-Johnen-Strasse
Jülich, 52425
Germany
Phone: 49 1632874330
Fax: 49 2461611880
Email: juergen.dukart@gmail.com

Published in: Front. Public Health, 19 November 2021
Sec. Digital Public Health
Volume 9 - 2021 | https://doi.org/10.3389/fpubh.2021.763621

Own contributions

Writing the manuscript, preparing figures, developing smartphone applications, contributing to the design of the experiment, writing analysis code, statistical data analysis, and contributing to the interpretation of results. Total contribution 80%

frontiers in Public Health

METHODS
published: 19 November 2021
doi: 10.3389/fpubh.2021.763621

JTrack: A Digital Biomarker Platform for Remote Monitoring of Daily-Life Behaviour in Health and Disease

Mehran Sahandi Far[1,2], *Michael Stolz*[1], *Jona M. Fischer*[1], *Simon B. Eickhoff*[1,2] *and Juergen Dukart*[1,2]*

[1] *Research Centre Jülich, Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Jülich, Germany,* [2] *Medical Faculty, Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany*

Health-related data being collected by smartphones offer a promising complementary approach to in-clinic assessments. Despite recent contributions, the trade-off between privacy, optimization, stability and research-grade data quality is not well met by existing platforms. Here we introduce the JTrack platform as a secure, reliable and extendable open-source solution for remote monitoring in daily-life and digital-phenotyping. JTrack is an open-source (released under open-source Apache 2.0 licenses) platform for remote assessment of digital biomarkers (DB) in neurological, psychiatric and other indications. JTrack is developed and maintained to comply with security, privacy and the General Data Protection Regulation (GDPR) requirements. A wide range of anonymized measurements from motion-sensors, social and physical activities and geolocation information can be collected in either active or passive modes by using JTrack Android-based smartphone application. JTrack also provides an online study management dashboard to monitor data collection across studies. To facilitate scaling, reproducibility, data management and sharing we integrated DataLad as a data management infrastructure. Smartphone-based Digital Biomarker data may provide valuable insight into daily-life behaviour in health and disease. As illustrated using sample data, JTrack provides as an easy and reliable open-source solution for collection of such information.

Keywords: mobile toolkit, mobile sensing, remote monitoring, health science, biomarkers

OPEN ACCESS

Edited by:
Mike Conway,
The University of Utah, United States

Reviewed by:
Hsin-Yao Wang,
Linkou Chang Gung Memorial Hospital, Taiwan
Wijaya Martanto,
MOH Office for Healthcare Transformation, Singapore

****Correspondence:***
Juergen Dukart
juergen.dukart@gmail.com

Specialty section:
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

Received: *25 August 2021*
Accepted: *28 October 2021*
Published: *19 November 2021*

Citation:
Sahandi Far M, Stolz M, Fischer JM, Eickhoff SB and Dukart J (2021) JTrack: A Digital Biomarker Platform for Remote Monitoring of Daily-Life Behaviour in Health and Disease. Front. Public Health 9:763621. doi: 10.3389/fpubh.2021.763621

INTRODUCTION

Neurological and psychiatric diseases typically present with symptoms that are complex, atypical, fluctuant in disease progression, and display high variability between patients (1). Current diagnostic and efficacy evaluation methods often rely on in-clinic visits and subjective evaluation by patients, caregivers or clinicians. In-clinic evaluation methods are often costly, time-consuming and limited in their quality and quantity of observations (2). In addition, they are often prone to high inter- and intra-rater variability (3). The aforementioned drawbacks of traditional diagnosis methods may affect the diagnostic process especially in the early stage of the disease where there is a lag between the onset of the pathological process and the onset of symptoms (4).

Psychiatric and neurological diseases are typically long-term illnesses that cause significant fluctuations in symptoms over time. Therefore, recall and reporting biases are the key difficulties in evaluating respective diseases in episodic in-clinic visits. Remote monitoring of patients in their everyday-life using sensor-based at smart technologies is rapidly evolving and may assist clinicians in facilitating early diagnosis and evaluating and adjusting interventions. There has been an evolving interest in using newly emerged smart sensor technologies for monitoring of patients (5–10).

Modern smartphones and wearables are equipped with various sensors including motion (i.e., acceleration, gyroscope), location [i.e., Global Positioning System (GPS)], environment (i.e., barometer, temperature, light) and health sensors (i.e., heart rate) (11, 12). This rich combination of sensors along with their ability to collect ecological momentary assessments (EMA), and information about social interaction (i.e., social media, messaging and phone calls) have made smartphones a potential alternative to in-clinic evaluation for various types of assessments (13, 14). Such health-related information being collected in clinical trials are often referred to as digital-biomarkers (DB) (15). DBs can provide objective, ecologically valid, and invaluable information for better understanding of specific diseases. In addition, DBs enable frequent assessments from larger target populations over longer periods of time and may thus provide detailed insight into inter- and intra-individual disease variability in daily-life (16).

Several contributions enabling the use of smartphones as an assessment tool have been recently introduced. The first set are commercial devices such as Fitbit[1], Garmin[2], Apple[3] and Samsung[4] devices. The main focus of these applications is to provide feedback on the daily activity of users by visualizing and showing notification regarding their heart rate, number of steps and kind of activity. However, most of these devices provide limited access to the raw data and do not support high-frequency data collection. A second type are applications and platforms developed by researchers such as AWARE (17), RADAR-base (18), Beiwe (19), mCerebrum (20), mPower (21) and many others. The main focus of these mostly open-source platforms is to enable data collection for research applications as well as to facilitate data sharing and reproducibility. Yet, these software packages are often limited by an often narrow focus to some specific clinical indications or with respect to privacy aspects. Also, these once in a while updated platforms make some of them unstable for the rapidly growing smartphone ecosystem.

Whilst there are several platforms that are able to collect context-driven data, the trade-off between privacy, optimization, stability and research-grade data quality is not well met yet. Thus, we aim to fill this gap by introducing the JTrack platform. JTrack was developed as an Android-based application for smartphones and an online server-side dashboard. JTrack application comprises the following main categories of components: sensor data, location data, Human Activity Recognition (HAR), and smartphone and application usage monitoring. Each component has the option to be used for active (with user interaction) and passive (without user interaction) monitoring. The dashboard side is an online platform to create and manage studies, integrating DataLad (22) infrastructures to facilitate management and sharing of collected data. JTrack is a modular open-source with a high level of optimization, security and privacy making JTrack a practical solution for clinicians and researchers to collect, manage, and share digital biomarker data.

METHODS

General Description

Here we introduce the main components of the JTrack platform (**Figure 1**) comprised of the JTrack app (**Figure 2**) and an online dashboard interface (**Figure 3**). The smartphone application JTrack was developed for smartphones with the Android operation system (OS). The reason for selecting Android was a wider range of users[5] (73%) and fewer restrictions which were necessary for technical aspects of application development.

JTrack enables passive 7/24 data collection running in the background. Active data collection is enabled through simple interaction (i.e., start and stop recording, i.e., before and after execution of a specific task). All collected data are recorded locally in the application and then synchronized on a periodic basis (i.e., connection to the Internet, have enough battery charge). All local data are deleted from the phone storage upon successful data transfer. To minimize the risk of data loss, we implemented auto-start functionality (without user interaction) to resist unwanted application crashes or operating system reboots, and all the crashes are reported via the Firebase dashboard[6].

On the server side, the JTrack dashboard was developed as an online web-application where study owners can create and manage studies. The dashboard consists of a front-end interface and a back-end API which is integrated with DataLad (22) as a data-management tool. The dashboard provides an overview of received data including sanity checks such as MD5 for received data, and embedded validity checking methods.

QR-Code Authentication

To provide a convenient and secure way of activation we implemented a QR-code method. The QR-code for each subject is generated as a pdf file from the dashboard. Each QR-code contains all the necessary information such as user ID, Study ID, and address of the target server or an optional authentication method (e.g., OAuth2). To join a study, the one-time QR-code needs to be scanned using the QR-scanner embedded in the JTrack app. Additional backup QR-codes are provided for scenarios in which users may want to leave and re-join or need to switch their device.

Location Service

Location service provides an update on visited location data such as longitude, latitude, and altitude. This service operates as a part of the passive recording. The location data can be inquired based on pre-defined periods (i.e., 10 min). To ensure anonymization, for each user, a random value is generated during activation on the phone, which shifts the latitude to a random place on the

[1] https://www.fitbit.com/global/us/home (accessed November 23, 2020).

[2] https://www.garmin.com/en-US/ (accessed November 23, 2020).

[3] https://www.apple.com/ (accessed November 23, 2020).

[4] https://www.samsung.com/us/ (accessed November 23, 2020).

[5] https://gs.statcounter.com/os-market-share/mobile/worldwide (accessed August 19, 2020).

[6] https://firebase.google.com (accessed November 23, 2020).

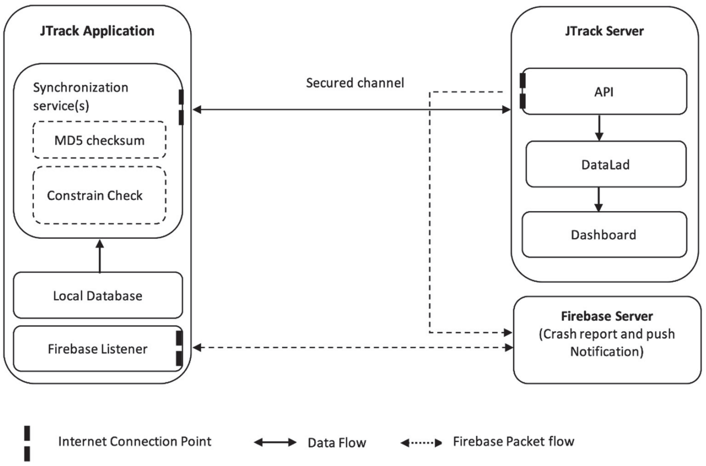


FIGURE 1 | JTrack Platform overview.

globe. In addition, all recorded coordinates are rotated using a randomly generated fixed degree around this initial coordinate to ensure that even if the true installation location is known no other coordinates can be derived. These values remain on the phone throughout the study and are deleted upon deinstallation of the app. To keep a high accuracy, each data point is first transferred from WGS-84 to Cartesian coordinates. After the transformation using the generated value, the coordinates are transferred back to their native space. Since this transformation occurs before actual recording, all the collected data is relative and cannot be used to recover the user's actual location. Furthermore, we used a fused method that provides more accurate data (median accuracy of 14 meters) by combining GPS and network information.

Human Activation Recognition

Inertial Measurement Unit (IMU) sensors of smartphones or wearables can be used to differentiate between human activities. Several studies described reliable algorithms for HAR (23, 24). Nowadays, these algorithms are routinely deployed in commercial devices, as well as in a wide range of research areas from medicine to military. JTrack uses the Google Play Activity Recognition Service[7] for HAR, which recognizes up to six types of activity (walking, running, still, on bicycle, on vehicle or tilting). JTrack can record the detected activity and the assigned certainty with a pre-specified frequency of 5 min. The HAR module is computationally lightweight, optimized and does not require direct access to raw sensor data.

Application Usage Statistic

JTrack can collect the statistic of user's interactions with the smartphone. This data includes the name of the application and the amount of time it is used in the foreground since the previous midnight. Phone calls and SMS are treated as applications with same usage statistics being collected as above. No content of the applications, messages or phone calls (including phone numbers) is collected at any stage.

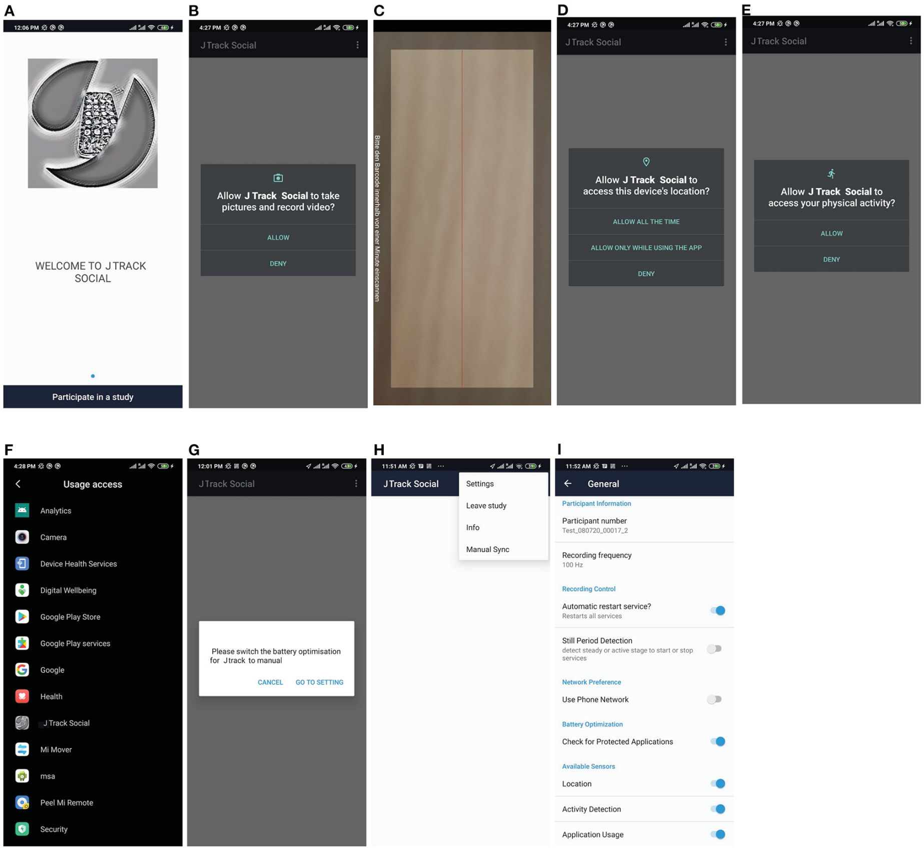
Sensors

Various sensors are embedded in any modern smartphones which are classified as hardware implementation (i.e., accelerometer, gyroscope, barometer) or software implementation (i.e., rotation sensor). JTrack enables collection of data from most of the available sensors depending on the device model and version of Android. Among these, accelerometer, gyroscope and gravity sensors are the most important sensors for researchers focusing on motion analysis (6, 25–29). As a default, JTrack provides recording of accelerometer and gyroscope data in the passive collection mode. Other sensors can be added upon the researcher's choice by using the provided template module which requires minimal coding effort. For each sensor, sampling frequency in Hz can be adjusted using the dashboard when creating a new study.

Dashboard

When creating a new study in the dashboard, all aspects of a study such as study name, duration, number of subjects, recording frequency, and categories of data to be collected can be customized. After creating a study, the dashboard will generate QR-codes which are used for enrolment into the study.

[7] https://developers.google.com/location-context/activity-recognition (accessed June 23, 2020).



WELCOME TO J TRACK SOCIAL

Allow **J Track Social** to take pictures and record video?

Bitte den Barcode innerhalb von einer Minute einscannen

Allow **J Track Social** to access this device's location?

Allow **J Track Social** to access your physical activity?

Please switch the battery optimisation for Jtrack to manual

Participant Information

Participant number: Test_080720_00017_2

Recording frequency: 100 Hz

Recording Control

Automatic restart service? Restarts all services ●

Still Period Detection: detect steady or active stage to start or stop services ○

Network Preference

Use Phone Network ○

Battery Optimization

Check for Protected Applications ●

Available Sensors

Location ●

Activity Detection ●

Application Usage ●

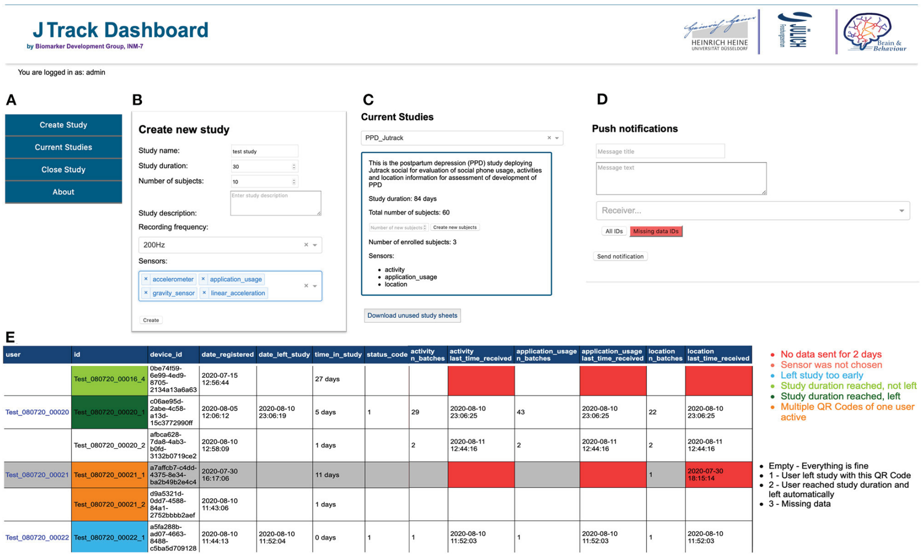
FIGURE 2 | JTrack Application Environment. **(A)** opening page of the application, **(B)** requesting for camera permission, **(C)** QR-Code scanner, **(D)** requesting for location permission, **(E)** requesting for an activity detection permission, **(F)** referring to ask for usage permission which is used for application usage module, **(G)** detecting a custom optimization and asking for disabling it, **(H)** the main menu of the application, where users may access the administration menu, leave a study, get information about the application and do manual synchronization, **(I)** administration menu, here we have access to the main setting of application, the information provided here is for further administration from study owners and most of the information is catch from sever during Installation.

The dashboard also provides management tools on an ongoing study producing information such as a number and time of received data for each sensor/modality and status (i.e., active, not active) of each participant in a particular study. We also implemented several quality controls including highlighting of missing data.

Furthermore, to assist study managers to establish further interaction with participants, we embedded a messaging method in the dashboard which allows to send a push notification directly to the participants' phone, either by selecting specific subject numbers or all participants within a particular study. Layered design (backend, frontend and data management layer) also makes the Dashboard flexible and extendable for further interaction and integration with third-party applications.

Performance, Security and Privacy

At all stages of the development, attention was paid to security and privacy as a main priority. In this context, we designed the JTrack platform to comply with GDPR and Google Developer



You are logged in as: admin

Create new study

Study name: test study

Study duration: 30

Number of subjects: 10

Study description: Enter study description

Recording frequency: 200Hz

Sensors: accelerometer, application_usage, gravity_sensor, linear_acceleration

Current Studies

PPD_Jutrack

This is the postpartum depression (PPD) study deploying Jutrack social for evaluation of social phone usage, activities and location information for assessment of development of PPD

Study duration: 84 days

Total number of subjects: 60

Number of new subjects

Number of enrolled subjects: 3

Sensors:

- activity
- application_usage
- location

Push notifications

Message title

Message text

Receiver...

user	id	device_id	date_registered	date_left_study	time_in_study	status_code	activity n_batches	activity last_time_received	application_usage n_batches	application_usage last_time_received	location n_batches	location last_time_received
	Test_080720_00016_4	0be74f59-6e99-4ed9-8705-2134a13a6a63	2020-07-15 12:56:44		27 days			🟥		🟥		🟥
Test_080720_00020	Test_080720_00020_1	c06ae95d-2abe-4c58-a13d-15c3772990ff	2020-08-05 12:06:12	2020-08-10 23:06:19	5 days	1	29	2020-08-10 23:06:25	43	2020-08-10 23:06:25	22	2020-08-10 23:06:25
	Test_080720_00020_2	afbca628-7da8-4ab3-b0fd-3132b0719ce2	2020-08-10 12:58:09		1 days		2	2020-08-11 12:44:16	2	2020-08-11 12:44:16	2	2020-08-11 12:44:16
Test_080720_00021	Test_080720_00021_1	a7affcb7-c4dd-4375-8e34-ba2b49b2e4c4	2020-07-30 16:17:06		11 days			🟥		🟥	1	2020-07-30 18:15:14
	Test_080720_00021_2	d9a5321d-0dd7-4588-84a1-2752bbbb2aef	2020-08-10 11:43:06		1 days							
Test_080720_00022	Test_080720_00022_1	a5fa288b-ad07-4663-8488-c5ba5d709128	2020-08-10 11:44:13	2020-08-10 11:52:04	0 days	1	1	2020-08-10 11:52:03	1	2020-08-10 11:52:03	1	2020-08-10 11:52:03

- No data sent for 2 days
- Sensor was not chosen
- Left study too early
- Study duration reached, not left
- Study duration reached, left
- Multiple QR Codes of one user active

- Empty - Everything is fine
- 1 - User left study with this QR Code
- 2 - User reached study duration and left automatically
- 3 - Missing data

FIGURE 3 | JTrack Dashboard Environment. **(A)** main menu, **(B)** here a new study can be created by specifying its details such as duration of the study, number of users, and list of data categories to be collected, **(C)** currently ongoing studies and details of the selected study can be found here. Also, the generated QR-Code can be downloaded here, **(D)** to accomplish more interaction with users participated in a particular study, a message as a push notification can be sent to a target user(s), **(E)** details of received data, date of registration, date left, duration within in study and quality control by color-labeling for sent data for users in a selected study can be monitored here.

Policies[8]. No sensitive data such as name, phone number, phone contacts or actual location are recorded at any stage. All the collected data transferred via Hypertext Transfer Protocol Secure (HTTPS) protocol and checked for any inconsistency using MD5 sanity checksum.

Concerning patient privacy, all users using JTrack are provided with clear information on what is been recorded and why. Permission requests for each module need to be approved during installation and activation. All participants may also stop and leave a study at any time directly from the app. Also, remote configuration and one-step recording allow clinicians to gain optimum control over the collected data without the need to collect any identifying information.

To reduce battery and memory usage, we provided several built-in optimizations such as:

- Detecting still period of the phone to pause recordings.
- Delete locally stored data right after synchronization with the server.
- Scheduled synchronization based on predefined criteria such as access to a Wi-Fi connection.
- Detect and provide a possibility to bypass performance optimizations (i.e., battery and memory) policies of phone manufacturers introduced on-top of the Android OS.

To reduce data loss due to crashes or reboots, automatic restarting is implemented alongside with Firebase integration to obtain performance and crash reports. Information about phone manufacturer, model, and OS version are among the optional recorded data (not active by default), which can be used to analyse and handle cross-sensor variability.

Pilot Study

In a pilot experiment, we collected in a cohort of healthy volunteers ($N = 21$, age: 26.1 ± 6.9, 7 female) for 2 weeks on a daily basis passively recorded data for application usage, location and activity recognition aside with self-reported estimates for these parameters [for application usage: time spent (in minutes) with the phone: total, social media and messenger]; for location and activity: walking/running distance (in meters). To test for associations between passively recorded and self-reported measures, we computed Pearson correlations across all subjects and time-points to compare both types of measures (i.e., merging location and activity recognition co compute distance covered by foot).

[8] https://developer.android.com/distribute/play-policies (accessed December 2, 2020).

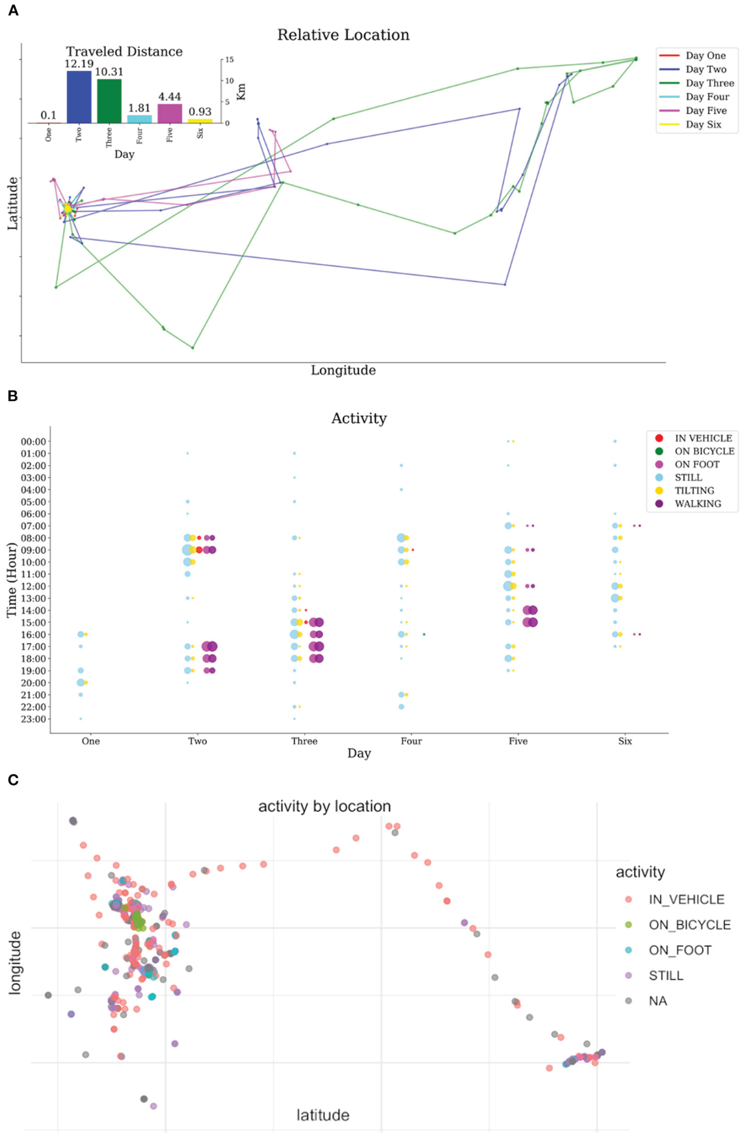


FIGURE 4 | Data sample for activity and location modules. **(A)** traveled distance and relative geolocation information for different days, **(B)** distribution of different physical activates during a different time of different days, **(C)** type of activity data combined with geolocation information.

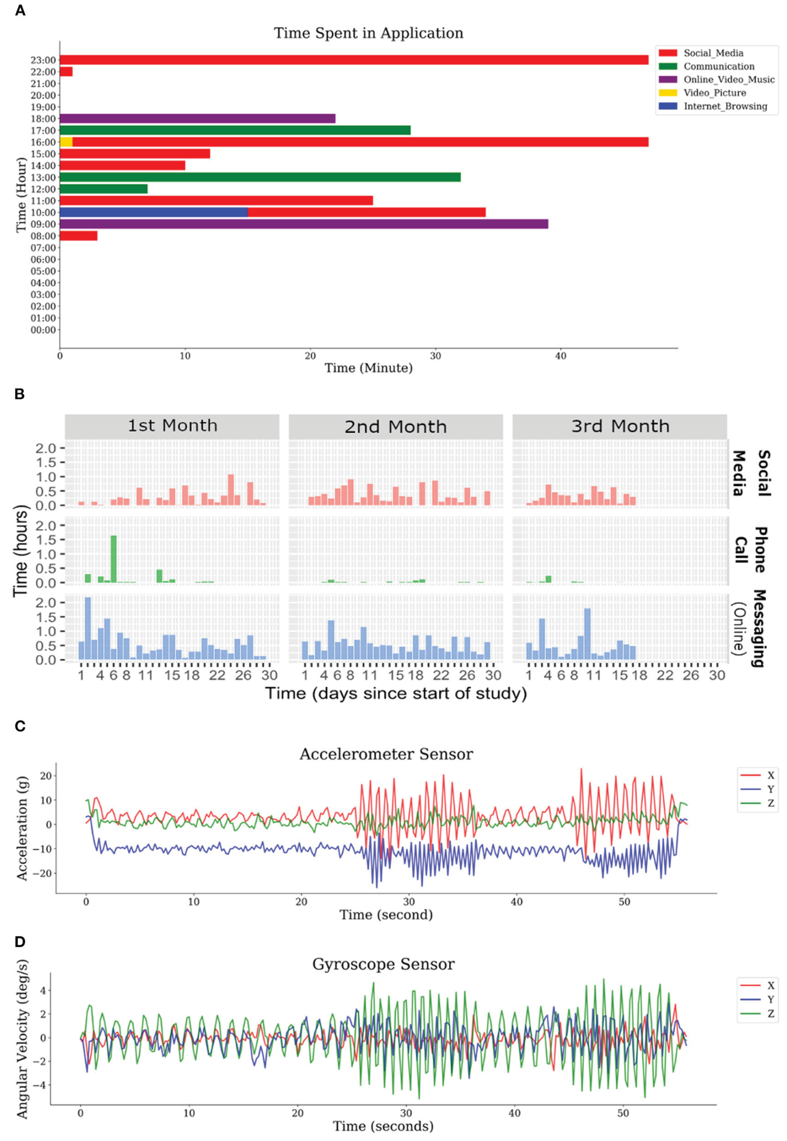


FIGURE 5 | Data sample for application usage module and sensors. **(A)** amount of time spent in different application types for a day, **(B)** distribution of different type of application usage during a study. **(C)** raw data recorded from the acceleration sensor, **(D)** raw data recorded from the Gyroscope sensor.

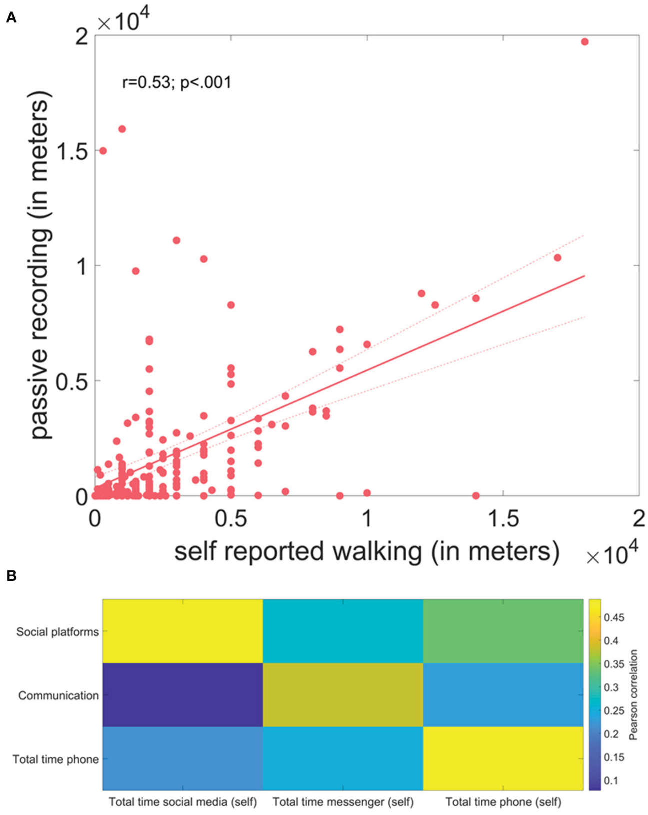


FIGURE 6 | Results of the pilot study comparing self-reported and passively recorded measures of daily behaviour. **(A)** Correlation plot between self-reported distance walking/running and distance estimates derived from fusion of location and activity recognition data. **(B)** Correlation matrix comparing self-reported measures of different smartphone usages and estimates derived from passively recorded app usage information.

RESULTS

To illustrate the utility of the JTrack application sample data were collected in the beta testing phase. **Figures 4**, **5** display such sample data collected for a single subject for different modalities. We provided sample data for each modality (i.e., location data, activity recognition, application usage and raw sensors data) also we further show a possible combination of the recorded data (i.e., location data with activity recognition data) in **Figure 4C**. Other combinations such as time and location (e.g., extracting amount of time spend outside of common residual place), location and application usage (e.g., extracting pattern of social interaction and applications used in-home condition) and activity and raw sensor data (e.g., extracting driving behaviour) are among the possible ways of making inference. **Figure 5B** also shows the daily phone and application usage (i.e., social media, phone calls, and online messaging platform) for a pilot participant.

In a pilot experiment, we further tested for associations between these passively recorded measures and self-reported estimates of specific behaviours. The self-reported distance covered by walking/running per day significantly correlated ($r = 0.53$; $p < 001$) with the information derived from passive monitoring (**Figure 6A**). Similarly, the time spent with the phone in total, communication and social platform significantly correlated with the passively-recorded estimates of respective phone usage measures ($r = 0.38$–0.49; all $p < 0.001$) (**Figure 6B**).

TABLE 1 | Comparison of existing frameworks with JTrack.

Framework	Location Anonymization	Official app Stores	Data Management	Remote Configuration	Activation	Customized OS Detection
AWARE (17)	NO	NO	SQL	YES	Text-based	NO
RADAR-base (18)	YES	YES	MongoDB	YES	QR-Code	NO
Beiwe (19)	NO	YES	PostgreSQL	NO	Text-based	NO
mCerebrum (20)	NO	NO	MySQL	NO	Text-based	NO
JTrack (this study)	YES	YES	DataLad	YES	QR-Code	YES

Comparison to Other Platforms

Table 1 shows how the JTrack platform compares to other similar and related platforms in terms of some key features such as security and privacy, activation, management and also stability. AWARE (17) is a platform for remote assessment of a wide range of phone sensors, activity and self-reported data. AWARE also supports additional plugins for external sensors and new data. However, this ability also requires a further declaration of permissions which limits control over privacy. mCerebrum (20) is another platform for remote assessment supporting a wide range of high-frequency sensors with a focus on energy-optimization. However, this platform has not been updated in while (latest update is May 2018 in their GitHub repository), questioning its performance on new versions of Android-OS. Beiwe (19) is the next platform supporting remote monitoring and DBs assessments which has a flexible study portal, modeling and data analysis tools. Nevertheless, this platform does not have local data storage and makes use of Amazon Web Services (AWS) cloud computing infrastructure. Such public cloud-based solutions are often more cost effective and convenient to use since as they simplify the build and maintenance process (this is particularly evident when the number of users and the data collected are small to medium sized), yet they may also raise data privacy questions and require additional deployment procedures. Another drawback of this platform is the collection of identifiable data such as phone number, media access control (MAC) address of WIFI and Bluetooth devices. RADAR-base (18) is the last open-source platform in the list. It has a well-organized structure which is using Confluent and Apache Kafka services and flexible study portal. Nevertheless, the deployment and adaptation of this platform require heavy configuration. Concerning the convenient registration, it requires text-based registration. Location data being collected in background is considered as a big concern in terms of privacy which is also frequently regulated by Google Developer Policies[9] and restricted by recent updates in Android OS. Among all the compared platforms only RADAR-base provides relative location. There are several alternative variants of these platforms that may strengthen some of the basic capabilities, such as Health Outcomes through Positive Engagement and Self-Empowerment (HOPES) which is based on the Beiwe platform (30) and AWARE-Light which is based on the aware framework. While these additional enhancements may address some of the shortcomings of the specific underlying platform, here we only focused on the comparisons to the core versions.

Easy one-step registration and authentication via QR-Code, as well as remote configuration, make JTrack more practical in both the usage and management aspects. Battery and memory optimizations offered by Android OS or phone manufacturers can affect the stability and consistency of data collected, JTrack provides built-in detection and circumvention methods for better stability that are not provided by comparator platforms at this level.

DISCUSSION

We developed JTrack as an open-source, smartphone-based platform for digital phenotyping. JTrack consists of a smartphone application and an online dashboard enabling remote data collection and study management. JTrack provides a flexible and modular environment for collection of various types of sensor and smartphone usage data with particular attention being paid to patient privacy as well as compliance with GDPR regulations.

From the functionality perspective, most of the solutions described above were developed with the focus on specific applications, i.e., a specific disease [i.e., RADAR-base (18) and Beiwe (19)]. Their application is therefore limited to the respective primary context. In contrast, some other platforms were developed to collect as much information as possible with little attention to data privacy [i.e., AWARE (17)]. Such frameworks violate GDPR and Google Play Store policies limiting their deployment for many clinical applications. JTrack aims to fill this gap by providing a customizable platform that can be deployed across different indications whilst paying large attention to privacy and security policies. JTrack aims to comply with GDPR regulations as well as with the Google Play Store policies. It only requires minimal access to the device information and avoids collection of identifiable or sensitive data.

Developing an application for smartphones always requires dealing with variation in devices (e.g., manufacture, screen size, available sensors) as well as the variation of operation systems (OS) versions. Different manufacturers may add further OS optimizations such as limiting background processes. This may cause inconsistencies in performance of monitoring applications. We introduced several layers to detect, report and prevent the side effects of these variations. JTrack is actively maintained and covers up to 84.9% of Android smartphones

[9] https://developer.android.com/distribute/play-policies (accessed December 2, 2020).

[Minimum Software Development Kit (SDK) 23] dealing currently with Android optimizations from eight main Android smartphone manufactures. Although the JTrack platform is now only available for the Android environment, which may introduce selection bias and limit participants to having an Android smartphone, an iOS version of JTrack is currently in development, with similar capabilities and will be made publicly available in the same GitHub repository and under the same open-source license.

Potential applications for JTrack include but are not limited to monitoring of motion information in diseases associated with alterations of gait and other motor functions affecting phone use. Similarly, the ability to track phone usage allows for monitoring of different types of behaviour, i.e., phone-based social interaction. As such, JTrack may be useful to track such behaviours in healthy participants as well its alterations by specific disorders.

Finally, to facilitate the reusability, JTrack is released under open-source Apache 2.0 licenses. All modules including online-management dashboard can be adopted and extended. It has been designed with modular structure to enable flexibility and customization to support new data and sensors.

Variations in device model, Android version, network quality, and other technical features may have negative effects on the performance of JTrack. Despite the effort to minimize crashes and data loss, there is no guarantee for such. During the development process, we used different third-party services (e.g., Google Play Service), any change or deprecation in these services, or Android policies may also affect the functionalities of JTrack partly or as a whole. Lastly, JTrack was designed and tested for smartphones. It may be used on other devices such as wearables (i.e., smartwatches) or tablets but further tests should be considered beforehand.

JTrack is an active and open-source project which is continuously maintained. We consistently improve and add new features to the platform. The features described here are part of the v1 release. Newer versions may differ and include additional functionalities at the time this article is published.

DATA AVAILABILITY STATEMENT

For the most updated and previous versions please visit the public repository accessible at https://github.com/Biomarker-Development-at-INM7. The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethikkommission an der Medizinischen Fakultät, Heinrich-Heine-Universität Düsseldorf Moorenstraße 5 40225 Düsseldorf. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MSF wrote the android-based application and performed all analysis. MSF wrote the manuscript with support of JD. MS and JF wrote the dashboard online application. JD designed the overall study. SE contributed to the overall design of the study. All authors contributed to the article and approved the submitted version.

FUNDING

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).

REFERENCES

1. Lyketsos CG, Kozauer N, Rabins PV. Psychiatric manifestations of neurologic disease: where are we headed? *Dialogues Clin Neurosci.* (2007) 9:111–24. doi: 10.31887/DCNS.2007.9.2/clyketsos
2. Zhan A, Little MA, Harris DA, Abiola SO, Dorsey ER, Saria S, et al. High frequency remote monitoring of Parkinson's disease via smartphone: platform overview and medication response detection. *arXiv Preprint arXiv:1601.00960.* (2016).
3. Rovini E, Maremmani C, Cavallo F. How wearable sensors can support parkinson's disease diagnosis and treatment: a systematic review. *Front Neurosci.* (2017) 11:555. doi: 10.3389/fnins.2017.00555
4. Balogh EP, Miller BT, Ball JR, Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, et al. *Improving Diagnosis in Health Care*. Washington, DC: National Academies Press (2015).
5. Rahlf AL, 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.* (2019) 1. doi: 10.3389/fspor.2019.00027
6. Orlowski K, Eckardt F, Herold F, Aye N, Edelmann-Nusser J, Witte K. Examination of the reliability of an inertial sensor-based gait analysis system. *Biomed Tech.* (2017) 62:615–622. doi: 10.1515/bmt-2016-0067
7. Hasegawa N, Shah VV, Carlson-Kuhta P, Nutt JG, Horak FB, Mancini M. How to select balance measures sensitive to Parkinson's disease from body-worn inertial sensors-separating the trees from the forest. *Sensors.* (2019) 19. doi: 10.3390/s19153320
8. 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) 2013:389195. doi: 10.1155/2013/389195
9. Cancela J, Pastorino M, Arredondo MT, Nikita KS, Villagra F, Pastor MA. Feasibility study of a wearable system based on a wireless body area network for gait assessment in Parkinson's disease patients. *Sensors.* (2014) 14:4618–33. doi: 10.3390/s140304618
10. Serra-Añó P, Pedrero-Sánchez JF, Hurtado-Abellán J, Inglés M, Espí-López GV, López-Pascual J. Mobility assessment in people with Alzheimer disease using smartphone sensors. *J Neuroeng Rehabil.* (2019) 16:103. doi: 10.1186/s12984-019-0576-y
11. Bandodkar AJ, Wang J. Non-invasive wearable electrochemical sensors: a review. *Trends Biotechnol.* (2014) 32:363–71. doi: 10.1016/j.tibtech.2014.04.005
12. Khan Y, Ostfeld AE, Lochner CM, Pierre A, Arias AC. Monitoring of vital signs with flexible and wearable medical devices. *Adv Mater Weinheim.* (2016) 28:4373–95. doi: 10.1002/adma.201504366

13. Kraft R, Schlee W, Stach M, Reichert M, Langguth B, Baumeister H, et al. Combining mobile crowdsensing and ecological momentary assessments in the healthcare domain. *Front Neurosci.* (2020) 14:164. doi: 10.3389/fnins.2020.00164
14. Harari GM, Lane ND, Wang R, Crosier BS, Campbell AT, Gosling SD. Using smartphones to collect behavioural data in psychological science: opportunities, practical considerations, and challenges. *Perspect Psychol Sci.* (2016) 11:838–54. doi: 10.1177/1745691616650285
15. Insel TR. Digital phenotyping: technology for a new science of behaviour. *JAMA.* (2017) 318:1215–6. doi: 10.1001/jama.2017.11295
16. Swan M. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. *J Med Internet Res.* (2012) 14:e46. doi: 10.2196/jmir.1988
17. Ferreira D, Kostakos V, Dey AK. AWARE: mobile context instrumentation framework. *Front ICT.* (2015) 2:1–9. doi: 10.3389/fict.2015.00006
18. Ranjan Y, Rashid Z, Stewart C, Kerz M, Begale M, Verbeeck D, et al. RADAR-base: an open source mHealth platform for collecting, monitoring and analyzing data using sensors, wearables, and mobile devices. *JMIR Mhealth Uhealth.* (2018) 7:e11734. doi: 10.2196/preprints.11734
19. Torous J, Kiang MV, Lorme J, Onnela J-P. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Ment Health.* (2016) 3:e16. doi: 10.2196/mental.5165
20. Hossain SM, Hnat T, Saleheen N, Nasrin NJ, Noor J, Ho J, et al. mCerebrum: a mobile sensing software platform for development and validation of digital biomarkers and interventions. *Proc Int Conf Embed Netw Sens Syst.* (2017) 2017. doi: 10.1145/3131672.3131694
21. Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using researchkit. *Sci Data.* (2016) 3:160011. doi: 10.1038/sdata.2016.11
22. Halchenko YO, Meyer K, Poldrack B, Solanky DS, Wagner AS, Gors J, et al. DataLad: distributed system for joint management of code, data, and their relationship. *J Open Sour Softw.* (2021) 6:3262. doi: 10.21105/joss.03262
23. Littman ML. Activity recognition from accelerometer data. *In Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence.* (2005). p. 1541–6.
24. Wan S, Qi L, Xu X, Tong C, Gu Z. Deep learning models for real-time human activity recognition with smartphones. *Mobile Netw Appl.* (2020) 25:743–55. doi: 10.1007/s11036-019-01445-x
25. Cho Y-S, Jang S-H, Cho J-S, Kim M-J, Lee HD, Lee SY, et al. Evaluation of validity and reliability of inertial measurement unit-based gait analysis systems. *Ann Rehabil Med.* (2018) 42:872–83. doi: 10.5535/arm.2018.42.6.872
26. Mundt M, Koeppe A, David S, Witter T, Bamer F, Potthast W, et al. Estimation of gait mechanics based on simulated and measured IMU data using an artificial neural network. *Front Bioeng Biotechnol.* (2020) 8:41. doi: 10.3389/fbioe.2020.00041
27. Schlachetzki JCM, Barth J, Marxreiter F, Gossler J, Kohl Z, Reinfelder S, et al. Wearable sensors objectively measure gait parameters in Parkinson's disease. *PLoS ONE.* (2017) 12:e0183989. doi: 10.1371/journal.pone.0183989
28. Moore ST, Yungher DA, Morris TR, Dilda V, MacDougall HG, Shine JM, et al. Autonomous identification of freezing of gait in Parkinson's disease from lower-body segmental accelerometry. *J Neuroeng Rehabil.* (2013) 10:19. doi: 10.1186/1743-0003-10-19
29. Rodríguez-Martín D, Samà A, Pérez-López C, Català A, Cabestany J. Posture transition analysis with barometers: contribution to accelerometer-based algorithms. *Neural Comput Applic.* (2020) 32:335–49. doi: 10.1007/s00521-018-3759-8
30. Wang X, Vouk N, Heaukulani C, Buddhika T, Martanto W, Lee J, et al. HOPES: an integrative digital phenotyping platform for data collection, monitoring, machine learning. *J Med Internet Res.* (2021) 23:e23984. doi: 10.2196/23984

Conflict of Interest: JD is a former employee and received consultancy fees on another topic from F. Hoffmann-La Roche AG.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Paper2: System Comparison for Gait and Balance Monitoring Used for the Evaluation of a Home-Based Training

Clara Rentz[1], Mehran Sahandi Far[2,3], Maik Boltes[4], Alfons Schnitzler[5,6], Katrin Amunts[1,7], Juergen Dukart[2,3] and Martina Minnerop[1,5,6]

[1] Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, 52428 Jülich, Germany

[2]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, 52428 Jülich, Germany

[3] Institute of Systems Neuroscience, Medical Faculty, Heinrich-Heine University Duesseldorf, 40225 Duesseldorf, Germany

[4] Institute for Advanced Simulation (IAS-7), Research Centre Jülich, 52428 Jülich, Germany

[5]Department of Neurology, Centre for Movement Disorders and Neuromodulation, Medical Faculty, Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany

[6]Institute of Clinical Neuroscience and Medical Psychology, Medical Faculty, Heinrich-Heine University Duesseldorf, 40225 Duesseldorf, Germany

[7]Institute for Brain Research, Medical Faculty, University Hospital Duesseldorf, Duesseldorf, Germany

Corresponding Author:

Clara Rentz
Heinrich-Heine University Düsseldorf, 40225

Published in: Sensors 2022, 22(13), 4975.
https://doi.org/10.3390/s22134975

Own contributions

Development of the Android-based application. Contributing to analysis and feature extraction from smartphone data and statistical data analysis and contributing to the interpretation of results. Total contribution 25%

sensors

MDPI

Article

System Comparison for Gait and Balance Monitoring Used for the Evaluation of a Home-Based Training

Clara Rentz [1,*], Mehran Sahandi Far [2,3], Maik Boltes [4], Alfons Schnitzler [5,6], Katrin Amunts [1,7], Juergen Dukart [2,3] and Martina Minnerop [1,5,6]

[1] Institute of Neuroscience and Medicine (INM-1), Research Centre Juelich, 52428 Juelich, Germany; k.amunts@fz-juelich.de (K.A.); m.minnerop@fz-juelich.de (M.M.)
[2] Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Juelich, 52428 Juelich, Germany; m.sahandi.far@fz-juelich.de (M.S.F.); juergen.dukart@gmail.com (J.D.)
[3] Institute of Systems Neuroscience, Medical Faculty, Heinrich-Heine University Duesseldorf, 40225 Duesseldorf, Germany
[4] Institute for Advanced Simulation (IAS-7), Research Centre Juelich, 52428 Juelich, Germany; m.boltes@fz-juelich.de
[5] Department of Neurology, Center for Movement Disorders and Neuromodulation, Medical Faculty, Heinrich-Heine University Duesseldorf, 40225 Duesseldorf, Germany; schnitza@med.uni-duesseldorf.de
[6] Institute of Clinical Neuroscience and Medical Psychology, Medical Faculty, Heinrich-Heine University Duesseldorf, 40225 Duesseldorf, Germany
[7] C. and O. Vogt Institute for Brain Research, Medical Faculty, University Hospital Duesseldorf, Heinrich-Heine University Duesseldorf, 40225 Duesseldorf, Germany
* Correspondence: c.rentz@fz-juelich.de

Abstract: There are currently no standard methods for evaluating gait and balance performance at home. Smartphones include acceleration sensors and may represent a promising and easily accessible tool for this purpose. We performed an interventional feasibility study and compared a smartphone-based approach with two standard gait analysis systems (force plate and motion capturing systems). Healthy adults (n = 25, 44.1 ± 18.4 years) completed two laboratory evaluations before and after a three-week gait and balance training at home. There was an excellent agreement between all systems for stride time and cadence during normal, tandem and backward gait, whereas correlations for gait velocity were lower. Balance variables of both standard systems were moderately intercorrelated across all stance tasks, but only few correlated with the corresponding smartphone measures. Significant differences over time were found for several force plate and mocap system-obtained gait variables of normal, backward and tandem gait. Changes in balance variables over time were more heterogeneous and not significant for any system. The smartphone seems to be a suitable method to measure cadence and stride time of different gait, but not balance, tasks in healthy adults. Additional optimizations in data evaluation and processing may further improve the agreement between the analysis systems.

Keywords: gait; balance; training; biomarkers; motion capturing; smartphone; IMU; video-based; home-based

check for updates

Citation: Rentz, C.; Far, M.S.; Boltes, M.; Schnitzler, A.; Amunts, K.; Dukart, J.; Minnerop, M. System Comparison for Gait and Balance Monitoring Used for the Evaluation of a Home-Based Training. *Sensors* **2022**, *22*, 4975. https://doi.org/10.3390/s22134975

Academic Editors: Winson Lee and Emre Sariyildiz

Received: 10 June 2022
Accepted: 28 June 2022
Published: 30 June 2022

1. Introduction

Gait and balance are impaired in aging, but also in various orthopedic and in particular neurological disorders. This impairment is often associated with reduced walking speed, increased gait variability or increased postural sway [1–3] and can lead to considerable constraints in daily life (e.g., bradykinesia/akinesia and freezing of gait in Parkinson's disease [4], unstable and wide-based gait in ataxias [5,6]). Identifying and assessing these constraints in daily life and providing suitable therapeutic (training) options such as physiotherapy is highly important.

There is an increasing demand to monitor physiological functions and disease-related symptoms independent of the physical presence of the respective participants or patients at the study site. Enabling study participation in a home-based setting, e.g., for human physiological monitoring [7,8] or by using wearables as measurement devices for assessing gait and balance [9,10], is an interesting and promising approach. Inertial measurement units (IMUs) consisting of accelerometers, gyrometers and magnetometers are routinely embedded in the hardware of smartphones. Due to their broad availability and the convenient option to implement applications, they may provide an attractive hands-on tool for measuring gait and balance in home-based settings. However, this set-up has been applied only recently in the field of motion analyses [10,11] and is not yet part of the standard clinical tools of measuring gait and balance.

Currently, the most commonly used instruments for gait and balance analysis are force plates (pressure-sensitive walkways) and body-worn motion capturing (mocap) systems based on IMU or optical data [12]. All these stationary systems allow the detection of abnormal or altered gait patterns in various neurological disorders such as Parkinson's disease (PD, [13]), multiple sclerosis (MS) or ataxias [14]. Using force plates (GAITRite, 5.1 m), it was shown that PD patients have a longer stride duration, a shorter stride length and greater variability in both, compared to healthy controls [13]. In addition, stride length and velocity were reduced in ataxia patients (force plate and body-worn sensors; [14]). These gait analysis systems were also able to detect performance changes after interventions. For example, Conradsson et al. [15] found improved gait velocity and stride length in normal gait after a ten-week balance training in PD patients. They measured normal walking with a GAITRite 9 m electronic walkway with and without a cognitive task and used the averaged result of six trials. Similarly, Giardini et al. [16] used the averaged results of four trials of normal walking at usual speed on a GAITRite 4.5 m electronic walkway and showed that two forms of physical exercise training (balance exercises and mobile platform training) improved gait speed in patients with PD, whereas only the balance exercises led to improved cadence and stride length.

Although the completion time of the Timed-Up-and-Go (TUG) test is used as a standard for quantifying functional mobility in a clinical context [17], electronic assessment of balance has increasingly been used in research [18]. The most commonly used instrument is a force plate (similar to gait analysis), however, an increasing number of technologies, whose reliability and validity was described in Baker et al. [19], are being used on a regular basis (e.g., inertial sensors). For balance tasks, center of mass or center of pressure data are commonly used to determine the area of postural sway, path length and mean velocity [20,21]. Morenilla et al. [22] described altered sway areas and velocities in PD patients when examining normal stance on a tri-axis force plate (Kistler). They found significant increases in total sway area and in mean anteroposterior and mediolateral displacement for PD patients. Moreover, Sun et al. [23] reported that both a new inertial body-worn sensor and a force plate were able to discriminate between subjects with severe MS and healthy control. However, only the force plate was able to distinguish subjects with mild MS from healthy control and patients with severe MS. Studies using force plates were also able to detect changes in performance after training interventions [24,25], i.e., patients with chronic stroke showed improved sway distance after participation in a virtual reality reflection therapy [26], and children with cerebral palsy showed decreased sway area and sway path after 12 weeks of training with a gaming balance board [27].

Thus, these stationary analysis systems of gait and balance are obviously able to detect performance differences between different groups in addition to shifts in performance over time or after intervention. They stand out in terms of their accuracy and ease of use. However, whether this also holds true for smartphone-based evaluation of gait and balance is still a topic of intensive research. In contrast to force plates and whole-body IMUs, the smartphone relies on a single sensor estimating velocity from acceleration and, in addition, gravitational influences and high-frequency noise must be filtered out. The advantage of smartphones would lie in their high disposability and saving of resources. Here we compared smartphone-based assessment of gait and balance tasks before and after three weeks of training to two commonly applied stationary gait analysis systems. We evaluated the feasibility of this approach to draw conclusions about the agreement of the three gait analysis systems and their ability to detect changes after a training intervention.

2. Materials and Methods

In this interventional feasibility study, smartphone-based evaluation of gait and balance was combined with two common stationary gait analysis systems requiring a laboratory environment: a zebris force plate and a Xsens mocap system with inertial sensors. Overall, 25 participants were recruited into the study. Two applications (apps, "*JTrack EMA*" and "*JTrack Social*") were installed on the smartphones of the participants (screenshots are available in Far et al. [28]). Both apps were developed at the Forschungszentrum Jülich [28]. *JTrack EMA* was developed for collection of ecological momentary assessments, so that common clinical questionnaires can be easily implemented into the app. *JTrack Social* was developed for customizable gathering of sensor data, including accelerometer information, using sensors embedded in any modern smartphone. Data were collected during a three-week video-based training intervention, which was performed at home and included twelve gait and balance training sessions, each lasting 20 min (see Figure 1). Participants were asked to indicate how many of the training videos they performed in total. Nevertheless, no verification of this information could take place. The present study was a feasibility study of a combined assessment and training protocol for gait and balance in healthy subjects. Written informed consent was obtained by all participants. The study was approved by the ethics committee of the Psychology faculty of the Heinrich Heine University Düsseldorf.

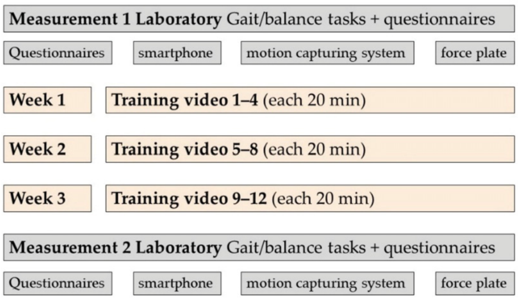

Measurement 1 Laboratory Gait/balance tasks + questionnaires
Questionnaires
smartphone
motion capturing system
force plate
Week 1
Training video 1–4 (each 20 min)
Week 2
Training video 5–8 (each 20 min)
Week 3
Training video 9–12 (each 20 min)
Measurement 2 Laboratory Gait/balance tasks + questionnaires
Questionnaires
smartphone
motion capturing system
force plate

Figure 1. Overview of study design.

2.1. Participants

Twenty-five participants were recruited via notices at universities, supermarkets and social media, and via newspaper. Participants had to be aged between 20 and 70 years, needed to walk safely without a walking aid, and did not report joint problems (osteoarthritis, endoprostheses) or other neurological, muscular or other medical problems affecting gait (e.g., falls, deep brain stimulation).

2.2. Gait Analysis Systems

The following three gait analysis systems (see also Figure 2) were used for assessment of gait and balance tasks in this study:

- The zebris FDM force plate (4.24 m, zebris Medical GmbH, Isny, Germany, https://www.zebris.de/en/medical/stand-analysis-roll-analysis-and-gait-analysis-for-the-practice, accessed on 30 June 2022) with the Noraxon® myoPressure software (Noraxon U.S.A., Inc., Scottsdale, AZ, USA, https://www.noraxon.com/our-products/myopres sure/, accessed on 30 June 2022). This uses capacitive pressure sensors to capture the pressure distribution in gait and balance.

- The Xsens mocap system consists of the MVN Awinda hardware and MVN Analyze software (Xsens Technologies B.V., Enschede, The Netherlands, https://www.xsens.com/motion-capture, accessed on 30 June 2022). It consists of 17 IMUs attached to each distinctive segment of the body fixed with body straps, which record angular velocity, acceleration, atmospheric pressure and the Earth's magnetic field with a frequency of 60 Hz.
- Individual Android-based smartphones of the participants on which the JTrack Social app was installed [28]. During all measurements, the accelerometer data of the smartphone were recorded using this app. The smartphone was placed in a waist bag.

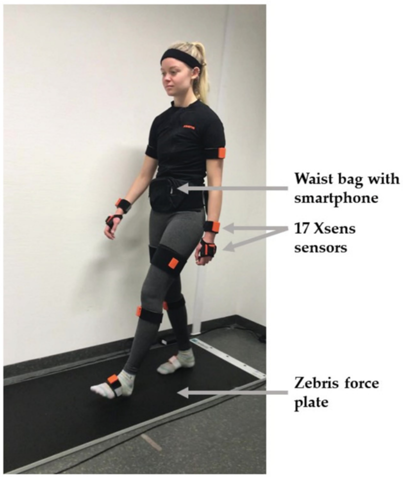


Figure 2. Representation of the three gait analysis systems used in the study.

2.2.1. Force Plate Feature Extraction

The zebris FDM *force plate* uses capacitive pressure sensors to capture the pressure distribution in gait and balance. No preprocessing was performed on the force and pressure data, which were recorded with a frequency of 100 Hz. Gait or balance reports are created automatically in the Noraxon myoPressure™ software, by selecting "Report" → "Bilateral Gait Report" for gait tasks and "Report" → "Stance Report" for stance tasks. The software uses the vertical ground reaction force to determine gait phases such as the heel strike or toe off. Movements in the beginning and at the end of the tasks that were not part of the task were unselected for all tasks. Apart from this, the entire distance walked on the force plate was included in the analysis. Feet positions were checked manually for tandem gait, since the software frequently was not able to distinguish the order of the left and right feet in this task. If foot positions were wrong according to the synchronized video, they were switched manually (left feet contacts were exchanged for right feet contacts).

In the report, stride time (s) describes the time between two heel contacts on the same side of the body. Cadence is the number of steps performed per second. The average velocity calculated for the force plate is the average stride length divided by the average stride time. Step width (cm) is the lateral distance between the center of the left and right heel.

2.2.2. Mocap System Feature Extraction

The Xsens *mocap system* computes the full-body motion based on constraints from a biomechanical model of the human skeleton with the help of sensor fusion algorithms. To configure the biomechanical model, body dimensions such as foot length, hip height and shoulder width of each participant were collated. The attached IMUs of the system are self-contained and light weight, so that they do not restrict subjects in their freedom of movement. After placing the system on a participant, a calibration process was performed as described in the MVN User Manual [29], i.e., to calculate the orientations of the sensors with respect to the corresponding segments. Quantities regarding the accuracy of the tracker and the MVN fusion engine can be found in the MVN User Manual [29]. A detailed description of the system is given in Schepers et al. [30].

The data were recorded with the Xsens MVN 2020.2 software and stored in the mvnx format after reprocessing in HD. A Python script was used to extract the position of the pelvis and both feet (foot segments located between the ankles within the Xsens model, see section 23.6.10 in the MVN User Manual [29]). The pelvis data were used to approximate the center of mass (COM, sensor position at the lower back on top of the sacrum). Data are given in the x-direction (anterior–posterior), in the y-direction (medial–lateral) and in the z-direction (vertical). The definition of axes also applies to the data of the left and right foot. The following procedures were separately repeated for each participant and each task.

Data were visualized to check for plausibility and to avoid including errors. Since the data contained turns at the end and at the beginning (most anterior and most posterior points, *x*-axis) of each lane, the first and last meters in the x-direction were excluded from the data. Data were then split into separate lanes (6 lanes for normal gait, 6 lanes for backward gait, 4 lanes for tandem gait) that every participant walked. IMU sensors showed a drift after a few lanes of walking, resulting in a mismatch between the correct direction of travel and the sensor-based detected direction of the *x*-axis as the main walking direction. This was corrected by rotating the data within the moving plane (x–y) to maximize the conformance between the walking direction and the *x*-axis. To calculate the time between two consecutive steps of the participant (step time), the vertical component of the COM data was used. As the COM moved up and down in cyclic movements, its peaks were used as markers for a step cycle. The height to find the peaks (scipy.signal, find_peaks) was adapted for each participant by visually checking the output plots. To avoid technical errors and enable single step detection during the tandem gait, an individual minimum distance between two consecutive peaks was required. The time between two steps (inter-step time) was calculated by subtracting the times of two neighboring peaks.

The step frequency (cadence), defined as the number of steps per second, is the inverse of the inter-step time.

Velocity as distance per time was calculated separately for each lane using the difference between the first and the last data point for position and time.

To calculate the lateral distance between both feet during steps (step width), the vertical *z*-axis and the *y*-axis (medial–lateral displacement) of the feet were considered. The time frame with the lowest foot position of each foot (mid-stance phase) was marked by searching for the minima in the z-direction (vertical axis). Its position in the y-direction at the same time frame was used to determine the distance between the left and right feet. Height and width in the find_peaks function were again adapted individually for each participant.

For the balance tasks, data import and inspection were performed in a similar way as described for the gait tasks. For each participant, the time span for analysis was selected in a way such that movements in the beginning or at the end of the balance task were excluded. Analysis was performed on the pelvis data (COM). The total path length that was traveled by the COM of the participant was calculated by summing the distance between all successive points in the path within the moving plane (x–y). The sway velocity described the number of millimeters the COM of the participant moved per second.

The area of an ellipse around the COM path was calculated by multiplying the anteroposterior sway and the mediolateral sway with pi.

2.2.3. Smartphone Feature Extraction

The JTrack Social app was installed on the individual Android-based smartphones of the participants and placed in a waist bag during the measurement (placed at the lower belly to approximate the COM while also ensuring simple handling).

All analyses of the JTrack Social app data were performed in MATLAB. The accelerometer data for each smartphone were recorded using the highest frequency provided for the respective smartphone (the recorded frequencies ranged between 100 and 252 Hz). All recorded gait and balance data were visually quality checked by removing non-tasks and, where identifiable, turn periods from the recordings. For normal gait data, manual step labeling was performed to obtain reference data for automated step labeling using a dedicated open-source MATLAB toolbox implemented for that purpose (https://github.com/juryxy/step_detector, accessed on 30 June 2022).

All accelerometer data were band-pass filtered in the range of 0.8–20 Hz to remove the gravitational component and the high frequency noise. Step detection for gait data was performed using the findpeaks function on the Euclidean norm of the accelerometer data. For this function, the following two parameters can be optimized for step detection—the minimum peak height (further expressed as standard deviation (SD) relative to the mean signal) and the minimum peak distance (in seconds). As the zebris FDM force plate was able to directly capture steps using pressure sensors, it was considered as closest to the ground truth together with the manually labeled data for normal gait. To identify optimum parameter combinations for smartphone step detection, we performed a grid search for the above parameters (peak height: 1.5 SD in steps of 0.1 to 3.0 SD; peak distance: 0.2 s in steps of 0.02 to 0.44 s), testing for correlations between the mean stride intervals (MSIs) obtained using these settings and MSIs derived using the ground truth provided by the force plate and manual labeling (Figure A1, Appendix A). For normal and backward gait, the optimum parameters providing the closest overall correlation to the ground truth were a minimum peak height of 2.3 SD and minimum peak distance of 0.38 s. For tandem gait, the optimum peak height was 2.7 SD and minimum peak distance was 0.42 s. Using these optimum parameters for step detection, the following features were computed using dedicated MATLAB scripts: stride time, cadence and velocity. To compute the mean velocity, we performed a step-wise double integration of accelerometer data to velocity and displacement using the first point as a reference. Thereby, the above band-pass filter was re-applied at each step to ensure that the residual gravitational and potential reintroduced high-frequency effects were removed from the data. Mean velocity (in m/s) was then computed as distance covered during the gait tasks divided by time.

For stance tasks, accelerometer data were transformed into displacement. The gravitational and high-frequency components were removed from acceleration and displacement data using band-pass filtering as for the gait tasks. Mean velocity was computed as point-by-point displacement divided by time. As the smartphone had no specific fixation of the phone orientation (except for a waist bag), the orientation of sensors with respect to the x- and y-plane differed across phones. To obtain an estimate of postural sway, we therefore performed a principal component analysis to determine the main directions of the sway in the three-dimensional space. The ellipsoid volume encompassing the 95% confidence interval of all points across the three principal components was computed as an estimate of postural sway around the COM (Figure A2, Appendix A).

An additional app, the JTrack EMA app (Biomarker Development, INM-7, Forschungszentrum Jülich), was used for the retrieval of questionnaires.

2.3. Study Tasks

2.3.1. Gait and Balance Tasks

For all gait tasks, participants were asked to walk safely across the force plate, then turn around behind the plate and walk back to the starting position. The walks were repeated several times with the number of iterations varying between tasks (for details see Table 1). For tandem gait, participants walked in a straight (imaginary) line by placing one foot in front of the other, placing the heel of one foot about a hand's width in front of the toes of the previous foot to enable separate foot detection by the force plate software. In the balance tasks, the participants were asked to keep their balance for as long as possible without leaving their position or holding up (maximum of 30 s). Participants performed all tasks without wearing shoes.

Table 1. Gait and balance tasks.

Task	Content
Normal gait (NG)	10 m × 4.24 m normal (forward) gait
Backward gait (BG)	6 m × 4.24 m backward gait
Tandem gait (TG)	4 m × 4.24 m tandem gait (walk on one line placing one foot in front of the other)
Narrow stance (NS)	Balancing in a narrow stance (feet close together)
Tandem stance (TS)	Balancing in a tandem stance (feet in one line)
Narrow stance with eyes closed (NSEc)	Balancing in a narrow stance with eyes closed
Single leg stance (SS)	Balancing on one leg

2.3.2. Questionnaires

Age, gender, body height, body weight, profession and years of education were retrieved in a demographic questionnaire during the first laboratory visit. To assess depression and anxiety, the German versions of the depression module of the patient health questionnaire (PHQ-9 [31], German version: [32]) and the hospital anxiety and depression scale ([33], German version: HADS-D [34]) were used. Additionally, general habitual well-being (FAHW [35]) and self-efficacy, optimism and pessimism (SWOP-K9 [36]) were assessed. To assess self-efficacy in relation to falls, the (modified) German version of the Activities-Specific Balance Confidence scale was used (ABC-D [37]).

The "PHQ_stress" and "PHQ_depression" subscores were selected from the PHQ-9 questionnaire. Although the depression and anxiety variables were used as exclusion criteria, the stress variable ranged from 0 to 20 and served as a covariate to describe the population. The anxiety subscore of the HADS-D had a cut-off value of >10 points and a depression subscore of >8 points. In the FAHW score, a total score of 38 to 50 or 35 to 47 (men and women, respectively) was defined as "average" according to the authors of the questionnaire. Additionally, the score contains a row of "smiley" icons, ranging from a happy face to a sad face. This was included in the evaluation by assigning a 1 to the happiest smiley and a 7 to the saddest smiley. The SWOP-K9 questionnaire contained items on self-efficacy (SWOP-SE), optimism (SWOP-OP) and pessimism (SWOP-PS), with scores ranging from 5 to 20, 2 to 8 and 2 to 8, respectively. For the ABC-D questionnaire, the scale was adapted to a 4-point response scale (not confident at all, somewhat less confident, somewhat confident, absolutely confident) so that a score between 16 (maximum confidence) and 64 (minimum confidence) could be achieved.

2.3.3. Training at Home

Gait and balance training was performed four times per week for 20 min by instruction via provided videos. The videos were produced by a physical therapy practice (PhysioStützpunkt, Köln, Germany) and uploaded to Vimeo (https://vimeo.com/, accessed on 30 June 2022). In each video, an experienced physiotherapist explained and demonstrated various tasks to improve gait and balance and instructed the participants to follow along. This included strength training, coordination training, stability training and mobility. The twelve videos progressed from simple to more demanding tasks and also included suggestions to reduce or increase the level of difficulty. Videos could be paused or repeated at any time, but participants were instructed to perform each training session only once until their second study visit was completed.

2.4. Statistical Analyses

From the set of extractable variables of each gait analysis system and each gait task, three variables were selected that were consistently available across all systems (see Table 2): *Gait velocity* (average velocity across all straight distances covered in the task, measured in meters per second), *stride time* (average duration of one stride defined as two consecutive steps in seconds) and *cadence* (average number of steps that are performed within one second). Additionally, *step width* was extracted from the force plate gait report and from the mocap system data, as this is an important variable to detect abnormal gait patterns (e.g., broadened base of support in cerebellar ataxias, see [3]). However, the step width cannot be derived from the acceleration data of the smartphone and was therefore not extracted from the smartphone data. For the balance tasks, the center of mass *(COM) sway area* (area of an ellipse enclosing all data points in the x- and y-direction) and the *velocity of the COM* (average distance in millimeters that the participant traveled per second) were chosen. These two variables showed good reliability in previous studies (e.g., [38,39]) and are commonly used for examining balance performance [20,21,40]. Both variables were available for all three gait analysis systems.

Table 2. Overview of gait and balance variables of all gait analysis systems used for statistical analysis.

	Output Variable	Description	Unit
Gait	Stride time	Time to complete one stride (two steps)	s
	Cadence	Number of steps per second	s^{-1}
	Velocity	Speed of movement	m/s
	Step width *	Lateral distance of left and right foot (center of heel) at one step	m
Balance	COM ellipse area (ellipsoid volume for smartphone)	Ellipse, enclosing 95% of all data points (100% in the mocap system) during a stance task (mediolateral and anteroposterior displacement)	mm^2 (mm^3)
	COM velocity	Speed of movement during a stance task (mediolateral and anteroposterior displacement)	mm/s

* not obtained with the smartphone.

Correlations between the questionnaire scores, between the individual variables *within one* gait analysis system, and between variables in *all* gait analysis systems, were calculated with the Pearson correlation coefficient. In this context, a correlation between 0.10 and 0.39 was described as weak, 0.40 to 0.69 as moderate and 0.70 to 1.00 as strong [40]. To analyze changes over time between the questionnaire scores and gait and balance variables at the first and second study visit (T1 and T2), either an ordinary paired-sample *t*-test was performed if the data scores were normally distributed, or a Wilcoxon rank test, if the data were not normally distributed. For all statistical analyses, a *p*-value of <0.05 was considered significant. Since results were corrected for multiple comparisons using a Bonferroni correction, the resulting *p*-values of <0.013 (force plate, mocap system) and <0.017 (smartphone) were considered significant when reporting changes over time. Boxplots of all gait and balance variables were checked and extreme outliers were excluded (>3 ∗ IQR above quartile 3).

3. Results

3.1. Participants

A total of 25 participants (age 44.0 ± 18.4 years) took part in the first study visit (T1, 52% female, 92% right-handed, see Table 3). One participant had missing data from the mocap system due to technical problems.

Table 3. Demographic information of all participants (n = 25). Education included school years plus years up to the highest graduation achieved (e.g., German Abitur equals 12 years of education). The HADS-D anxiety score had a cut-off value of >10 and the HADS-D depression score had a cut-off value of >8. The PHQ stress score had a maximum of 20 points.

	Mean ± SD	Range (Min.–Max.)
Age [years]	44.1 ± 18.4	20–71
Body height [cm]	172.3 ± 9.9	154–193
Body weight (n = 17) [kg]	67.6 ± 14.2	43–97
Education [years]	15.2 ± 3.2	10–25
HADS-D Anxiety [score]	3.3 ± 2.8	0–9
HADS-D Depression [score]	2.6 ± 2.6	0–10
PHQ Stress [score]	2.8 ± 2.1	0–8

For the second study visit, four participants dropped out (injury independent of the study (one), technical difficulties (one) and time constraints (two)). This led to a sample of 21 participants at T2 with an average age of 44.7 ± 19.4 years (57% female, 95% right-handed). All subjects reported having performed each of the training videos (12/12).

All demographic variables and questionnaire scores except the ABC-D score were normally distributed. Because one participant showed a depressive mood (HADS-depression score 10), all analyses were conducted with and without this subject. Since results did not differ, data from this participant were not excluded from further analyses.

Of the gait and balance variables, 8 of 33 gait variables were not normally distributed and 21 of 24 balance variables were not normally distributed. Accordingly, non-parametric statistical tests were selected for these variables. For detailed specifications of the variables, please see Table A2 (Appendix A).

3.2. Questionnaires

No differences between the questionnaires obtained at both study visits were found between T1 and T2 (Table 4, $p > 0.09$).

Table 4. Descriptive statistics of the questionnaire scores at the first and second study visit (T1, n = 25, and T2, n = 21). SE = self-efficacy (possible range: 5 to 20), OP = optimism (possible range: 2 to 8), PS = pessimism (possible range: 2 to 8). Activities-Specific Balance Confidence scale (ABC-D, possible range: 16 to 64), general habitual well-being (FAHW, average reference values between 35 and 50, smiley score ranging from 1 to 7).

	T1		T2	
Questionnaire [Score]	Mean ± SD	Range (Min.–Max.)	Mean ± SD	Range (Min.–Max.)
SWOP-SE	3.080 ± 0.49	2.0–3.8	3.229 ± 0.4485	2.2–4.0
SWOP-OP	3.240 ± 0.631	2.0–4.0	3.119 ± 0.7891	1.5–4.0
SWOP-PS	1.740 ± 0.614	1.0–3.0	1.667 ± 0.7130	1.0–3.0
ABC-D *	17.96 ± 2.574	16–28	17.76 ± 2.343	16–24
FAHW	59.12 ± 16.821	21–83	54.55 ± 25.310	−5–86
FAHW Smiley	2.04 ± 0.611	1–3	2.25 ± 0.786	1–4

* The ABC-D scores were not normally distributed. A Wilcoxon rank test was performed.

3.3. Gait and Balance Performance

3.3.1. Conformity of the Systems

Significant correlations between corresponding gait variables (stride time, cadence, velocity) across the three systems were present during all gait tasks. For the velocity variable during the backward and tandem gait, the correlations involving the smartphone were weak and did not all reach significance; correlations for the other two variables were significant.

For normal gait (Table 5), strong correlations were found between the three corresponding gait variables (stride time, cadence, velocity) of the force plate, mocap system and smartphone, except for one moderate correlation of velocity between the mocap system and smartphone. Step width was moderately correlated between the force plate and mocap system.

Table 5. Between-system correlations for normal gait between the force plate, mocap system and smartphone at T1 (first measurement time). Correlation after Pearson.

Normal Gait		**Force Plate** **(n = 23)**	**Mocap System** **(n = 22)**			**Force Plate** **(n = 24)**
Smartphone	Stride time	0.977 **	0.962 **	Mocap system	Stride time	0.981 **
	Cadence	0.942 **	0.934 **		Cadence	0.992 **
	Velocity	0.705 **	0.648 **		Velocity	0.925 **
	Step width				Step width	0.430 *

* Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed). n = number of participants included in the analysis.

For backward gait (Table 6), strong correlations were found between the stride time variables of all systems and for cadence between the force plate and smartphone. The remaining correlations regarding cadence and velocity were moderate or even showed no correlation for velocity between the mocap system and smartphone.

Table 6. Between-system correlations for backward gait between the force plate, mocap system and smartphone at T1 (first measurement time). Correlation after Pearson.

Backward Gait		**Force Plate** **(n = 23)**	**Mocap System** **(n = 22)**			**Force Plate** **(n = 24)**
Smartphone	Stride time	0.936 **	0.706 **	Mocap system	Stride time	0.731 **
	Cadence	0.919 **	0.685 **		Cadence	0.687 **
	Velocity	0.508 *	−0.019		Velocity	0.453 *
	Step width				Step width	0.361

* Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed). n = number of participants included in the analysis.

For tandem gait (Table 7), correlations were again strong between stride time and cadence variables across all three systems. However, for velocity, only moderate correlation was found between the force plate and the mocap system, but not between the smartphone and the two standard systems.

For balance tasks, moderate to strong significant correlations were found between the corresponding variables of the force plate and mocap system (see Table 8). For smartphone data, only three variables reached statistical significance (moderate correlations between the ellipse variables in tandem stance and the velocity variables in narrow stance with eyes closed between the force plate and smartphone, and a moderate correlation between the velocity variables in single leg stance between the mocap system and smartphone).

Table 7. Between-system correlations for *tandem* gait between the force plate, mocap system and smartphone at T1 (first measurement time). Correlation after Pearson.

	Tandem Gait	Force Plate (*n* = 17)	Mocap System (*n* = 19)			Force Plate (*n* = 19)
Smartphone	Stride time	0.875 **	0.899 **	Mocap system	Stride time	0.901 **
	Cadence	0.794 **	0.869 **		Cadence	0.861 **
	Velocity	0.149	0.365		Velocity	0.618 **
	Step width				Step width	−0.150

** Correlation is significant at the 0.01 level (2-tailed). *n* = number of participants included in the analysis.

Table 8. Between-system correlations for the stance tasks at T1. Cor. = correlation after Pearson. NS = narrow stance. TS = tandem stance. NSEc = narrow stance with eyes closed. SS = single leg stance. The number of participants included in each analysis varied between 14 and 24.

			Force Plate	Mocap System				Force Plate
Smartphone	Narrow stance	Ellipse	−0.072	0.093	Mocap system	Narrow stance	Ellipse	0.697 **
		Velocity	0.186	0.190			Velocity	0.673 **
	Tandem stance	Ellipse	0.550 *	0.315		Tandem stance	Ellipse	0.483 *
		Velocity	0.008	0.123			Velocity	0.468 *
	Narrow stance eyes closed	Ellipse	0.120	−0.058		Narrow stance eyes closed	Ellipse	0.782 **
		Velocity	0.580 *	0.210			Velocity	0.752 **
	Single leg stance	Ellipse	0.453	0.479		Single leg stance	Ellipse	0.672 **
		Velocity	0.243	0.528 *			Velocity	0.706 **

* Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).

3.3.2. Reference Values

To put the outcome values of the gait tasks in context, reference values from the literature are given in Table 9.

Table 9. Overview of values of gait variables found in the literature versus results of this study. A value description is given, unless values are mean ± SD.

		Literature			Own Results
		Values	System	Reference	(Force Plate, Mocap System, Smartphone)
Normal gait	stride time [s]	1.16 (0.92–1.41) (median (5th–95th percentiles))	zebris force plate	Pawik et al., 2021 [41]	1.18, 1.20 and 1.20
		1.09 ± 0.08	zebris force plate	Kasović et al., 2020 [42]	
	cadence [steps/s]	1.83 ± 0.17	zebris force plate	Kasović et al., 2020 [42]	1.66, 1.70 and 1.67
		1.72 ± 0.17	GAITRite force plate	Rao et al., 2011 [43]	
	velocity [m/s]	1.25 ± 0.14	zebris force plate	Kasović et al., 2020 [42]	0.98, 0.97 and 1.18
		0.94 ± 0.25	GAITRite force plate	Rao et al., 2011 [43]	
	step width [cm]	11.65 ± 2.85	zebris force plate	Kasović et al., 2020 [42]	11.64 and 10.6
		5–13 (usual walking base)		Whittle, 2007 [44]	
		11 ± 4	GAITRite force plate	Rao et al., 2011 [43]	

Table 9. *Cont.*

		Literature			**Own Results**
		Values	**System**	**Reference**	**(Force Plate, Mocap System, Smartphone)**
Backward gait	stride time [s]	1.2 ± 0.1	zebris force plate	Gimunová et al., 2021 [45]	1.22, 1.21 and 1.23
	cadence [steps/s]	1.68 ± 0.15	zebris force plate	Gimunová et al., 2021 [45]	1.66, 1.66 and 1.67
	velocity [m/s]	0.87 ± 0.12	zebris force plate	Gimunová et al., 2021 [45]	0.69, 0.66 and 0.55
		0.98 ± 0.23	GAITRite force plate	Edwards et al., 2020 [46]	
	step width [cm]	16.8 ± 4.87	zebris force plate	Gimunová et al., 2021 [45]	18.08 and 11.86
Tandem gait	cadence [steps/s]	0.8 ± 0.05 (estimated mean ± SD at 1 km/h speed)	zebris ultrasound system	Kronenbuerger et al., 2009 [47]	1.23, 1.19 and 1.23
		0.87 ± 0.29	GAITRite force plate	Rao et al., 2011 [43]	
	velocity [m/s]	0.27 ± 0.13	GAITRite force plate	Rao et al., 2011 [43]	0.45, 0.4 and 0.20
	step width [cm]	3.5 ± 2.6	GAITRite force plate	Rao et al., 2011 [43]	2.24 and 2.44

3.3.3. Differences over Time—Force Plate

Since not all variables were normally distributed, *p*-values either refer to *t*-tests (no indication) or to Wilcoxon-rank tests (indicated by "(W)").

For normal gait, a significant difference was found in all variables between T1 and T2: stride time (*p* = 0.003, Figure 3A), cadence (*p* = 0.002, Figure 3B), velocity (*p* = 0.002, Figure 4A) and step width (*p*(W) = 0.004, Figure 4B). For the backward gait, only the velocity variable (*p* = 0.005, Figure 4A) remained significant after correcting for multiple comparisons. For tandem gait, none of the variables remained significant after correcting for multiple comparisons.

For the stance tasks, none of the variables remained significant after correcting for multiple comparisons (Figures 5A and 4B).

The exact values for all tasks and gait analysis systems are reported in Table A1, Appendix A.

3.3.4. Differences over Time—Mocap System

In contrast to the force plate, a significant difference in normal gait was found in only two of four variables: stride time (*p* = 0.002, Figure 3C) and cadence (*p* = 0.001, Figure 3D). For the backward gait, only the velocity variable (*p* = 0.007, Figure 4C) remained significant after correcting for multiple comparisons—similar to the results of the force plate. For the tandem gait, a significant difference was found for two of four variables: for the stride time (*p* = 0.003, Figure 3C) and the cadence (*p* = 0.001, Figure 3D). No significant effect was found for the step width (Figure 4D); however, this may be related to the initial calibration procedure: the closer the participants' feet were in the "neutral position", the smaller the absolute values of the step width were in the later analysis.

Similar to the force plate, the mocap system analysis did not reveal a significant difference between T1 and T2 for any of the stance tasks.

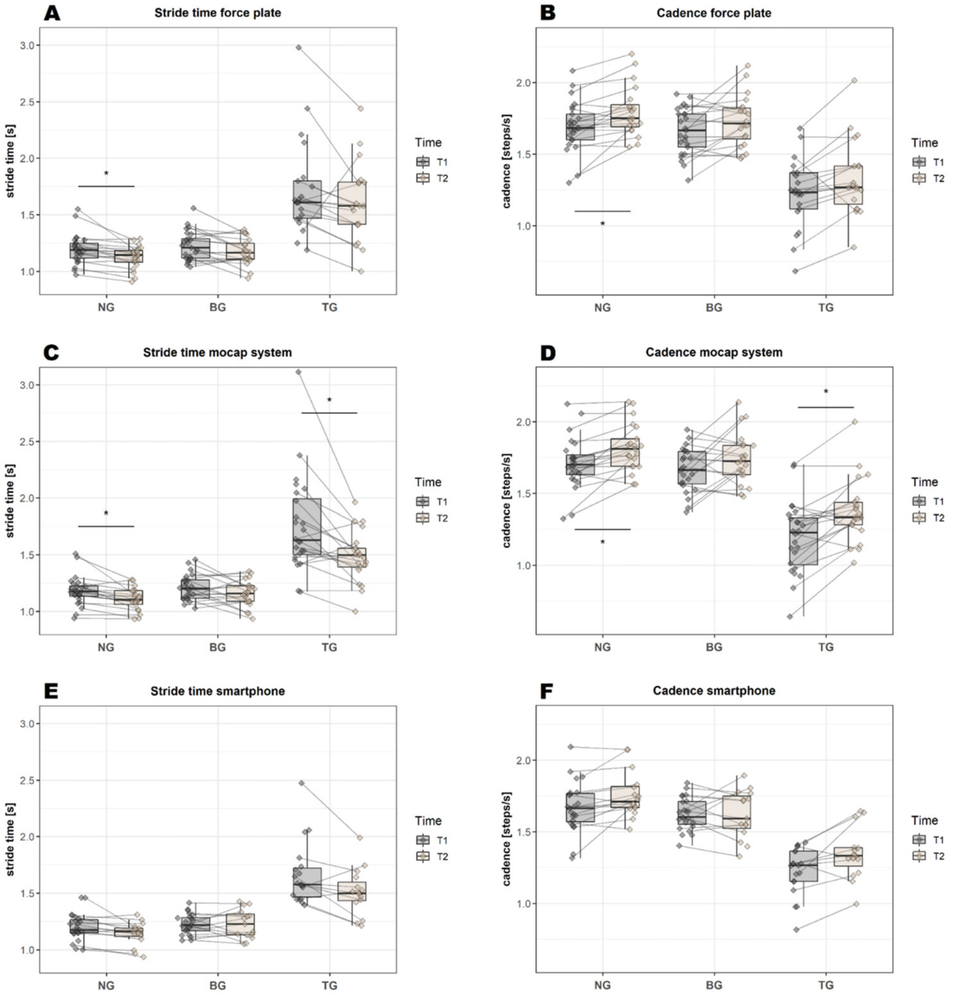


Figure 3. Graphical representation of the mean values of stride time and cadence for all three gait analysis systems at T1 and T2 (before and after training). Significant differences over time (after Bonferroni correction) are highlighted by an asterisk. BG = backward gait, NG = normal gait, TG = tandem gait.

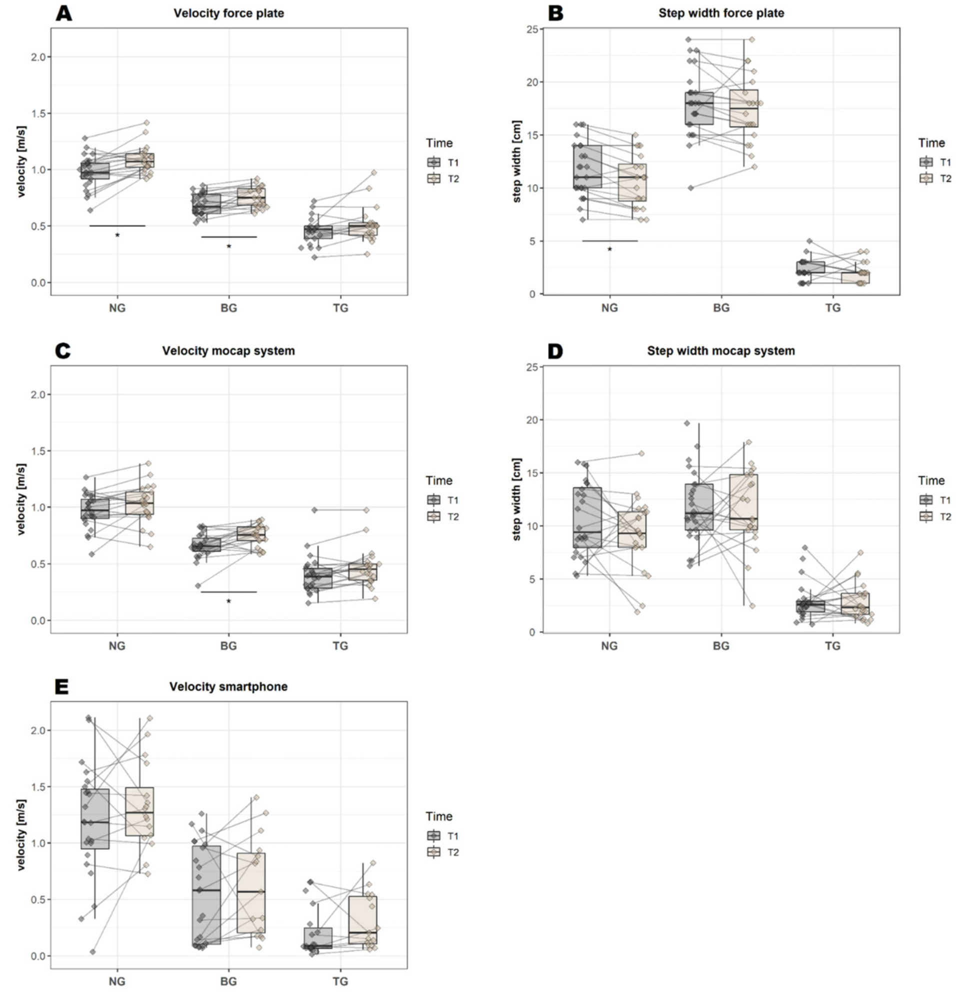


Figure 4. Graphical representation of the mean values of velocity and step width for all three gait analysis systems at T1 and T2 (before and after training). Significant differences over time (after Bonferroni correction) are highlighted by an asterisk. BG = backward gait, NG = normal gait, TG = tandem gait.

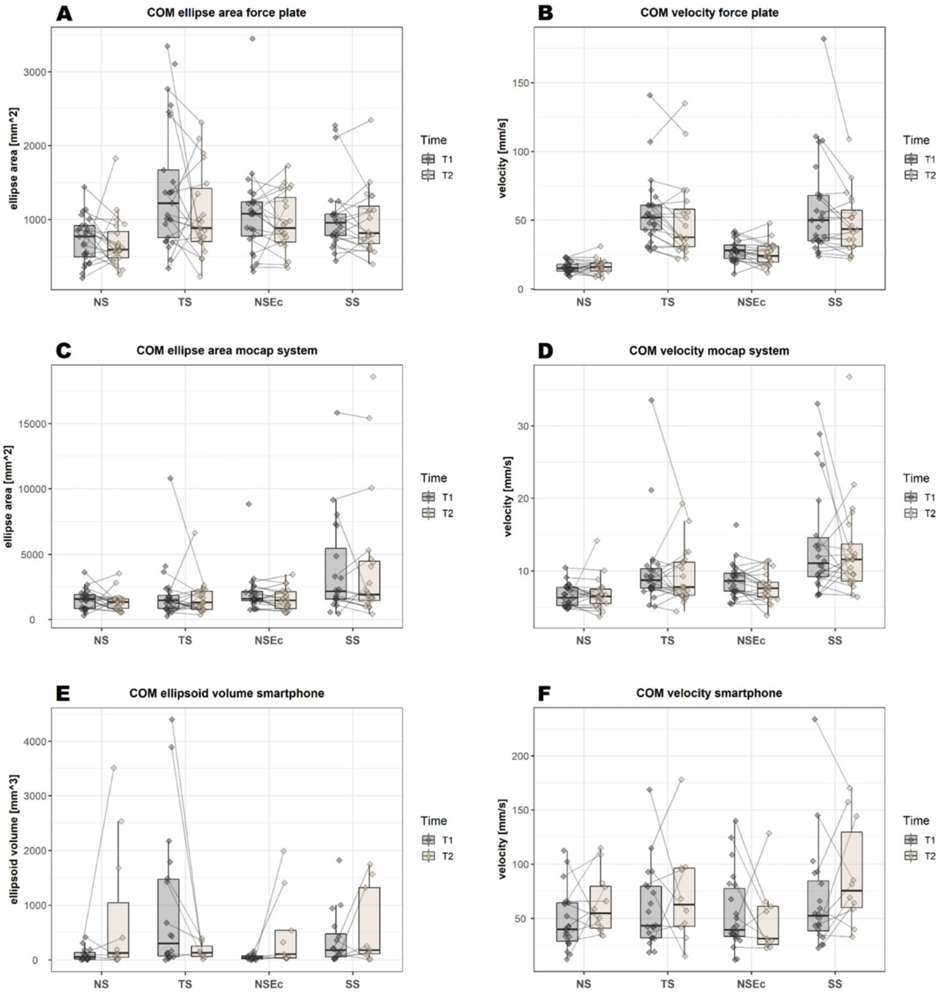


Figure 5. Graphical overview over the balance variables (center of mass ellipse area and velocity) in all three gait analysis systems at both measurement times (first measurement, T1, second measurement, T2). COM = center of mass, NS = narrow stance, TS = tandem stance, NSEc = narrow stance with eyes closed, SS = single leg stance.

3.3.5. Differences over Time—JTrack Smartphone Platform

In contrast to both the force plate and mocap systems, none of the variables of normal gait, backward gait or tandem gait remained significant after correcting for multiple comparisons (Figures 3E,F and 4E). Compared to the other gait analysis systems, the smartphone had a much higher variability of the velocity values, e.g., velocity values of the backward gait at T1 were 0.69 ± 0.09 m/s for the force plate and 0.55 ± 0.43 m/s for the smartphone (see Table A1, Appendix A).

Similar to both the force plate and mocap systems, the smartphone analysis showed no significant differences between T1 and T2 for any of the stance tasks (Figure 5E,F).

4. Discussion

Here, we performed an interventional feasibility study and compared three systems for the monitoring of home-based gait and balance training in healthy adults. In particular, we assessed the applicability of smartphone-based data collection in comparison to standard methods and the capability of the methods to detect performance changes after training.

4.1. Conformance of the Three Gait Analysis Systems

Gait variables obtained with both standard analysis systems (force plate and mocap) showed moderate to strong intercorrelations, except for step width. However, the strength varied depending on the performed gait task with excellent correlations for normal gait. Step detection during backward or tandem gait was more challenging and error-prone compared to normal gait, since feet were placed more cautiously and slowly, resulting in lower force and acceleration values, in addition to atypical movement patterns. In line with this, step width values correlated moderately between both systems for normal but not for backward gait. For tandem gait, the correlation between the step width values of both systems even revealed negative values, due to the calibration process of the mocap system [29]: if participants placed their feet in a very narrow stance during the "neutral position", required for the calibration process, the absolute values of the step width were much lower in the later analysis. This led to incorrect lateral positions of the feet and even to negative step width values in the tandem gait. For future studies using mocap systems, a standardized stance position of the participants is therefore highly recommended.

The JTrack based smartphone evaluation using accelerometer data showed strong correlations for the stride time and cadence variables of all gait tasks with both standard systems. Velocity, however, showed only moderate to strong correlations for normal and backward gait, and weak correlations for tandem gait. Taken together, all three gait analysis systems showed excellent agreement during normal gait, followed by the tandem gait task and a substantially lower agreement for the backward gait task. The agreement was better for the gait variables of stride time and cadence than for velocity. The less accurate velocity estimation via smartphone relied on a single sensor estimating velocity from acceleration using the first recorded value as a reference. As this first value was not calibrated in our study (i.e., no fixed position was taken of the phone when recording started), this may lead to biases in estimation of the initial velocity. It also explains the lack of correlation with other systems for tandem gait, for which the velocity was substantially lower, thereby increasing the impact of noise.

The strong correlations of smartphone-based gait variables with standard gait analysis systems found in our study are in contrast to Steins et al. [48], who described only moderate agreement between an iPod touch and an Xsens sensor when investigating the reliability of inertial sensors of smart devices during normal gait in healthy adults. Nevertheless, other studies suggested that smart devices are an acceptable method for assessing gait in rheumatic patients [49] and have the potential for future use in the clinic [13].

The stance variables of ellipse area and velocity showed moderate to strong correlations between the two standard force plate and mocap systems (see Section 3.3), in spite of large differences in the absolute values obtained with these methods (see Table A2, Appendix A). In contrast, only weak to moderate correlations were found between the smartphone and both other systems. This might be due to specific aspects of data acquisition and analysis. Force plates can directly register the foot print and determine the respective variables from position data. In contrast, the smartphone uses accelerometer information with respect to the first recorded value and thus only infers position data through double integration. Thereby, gravitational influences and high-frequency noise must be filtered out using band-pass filtering, which may lead to additional biases in position estimation. The mocap system uses multiple sensors, e.g., directly on the feet, and, in addition to the accelerometer data, also considers angular velocity, atmospheric pressure and magnetic field data, and a biomechanical model. This contrasts with the smartphone analyses, which relied on a single sensor near the COM. This enables the mocap system to determine the positions of the sensors relative to one another and to better estimate the gravitational and the noise components. Since the position and orientation of the smartphone were not fixed when recording started, the initial estimates may be biased, affecting all derived measures. Moreover, as the three axes in space were not fixed, it is difficult to determine an area in mm^2 in a standardized manner. Accordingly, the ellipse volume was computed in mm^3, introducing an additional source of variation.

Taken together, stride time and cadence seem to be variables that are robust to measurement with a smartphone, whereas other gait and stance variables are subject to some limitations.

4.2. Questionnaires

Since physical activity has a significant impact on mental well-being and vice versa, the objective motor assessment in this study was accompanied by a set of questionnaires addressing different aspects of subjective participant-reported outcome measures (e.g., depression- and anxiety-related symptoms, general well-being, stress, self-efficacy, optimism, pessimism and balance confidence).

Contrary to our expectations, the questionnaire scores did not differ between the pre- and post-training study visits. Physical therapy or exercises can reduce fatigue and improve one's emotional life [50] and mental health, in a manner that is even similar to psychotherapy. By comparison, our participants already had above-average FAHW scores at their first visit (reference values are given in [35]), indicating that the general well-being was already at a high level before the training and hence left less room for improvement.

Due to several constraints (study duration, compliance), a three-week period was chosen as the training interval in this study. Although Mikkelsen et al. [51] reported that exercising for 15 min three times per week already reduced depressive symptoms, most studies chose a longer time period for the training program or a longer duration for each unit to maximize the effectiveness of balance training and to prevent falls [52,53]. In the more specific context of home-based training, the highest effectiveness of video-based rehabilitation programs was found after at least four weeks [54]. Nevertheless, although a higher training volume or frequency can lead to better training results, it may also reduce compliance, as the subjective cost may exceed the perceived benefit of the training. In Haines et al. [55], a drop in compliance was found after three weeks. In our study, all subjects reported having performed each of the training videos, but verification of this information was not possible, impeding a valid statement regarding compliance.

4.3. Gait Performance

Mean values of stride time, cadence, velocity and step width obtained in our study were comparable to those found in the literature for normal gait in healthy adults (see Table 9). Similarly, stride time and cadence values during backward gait were comparable between the literature [45] and between all three gait analysis systems. However, in our study, velocity values were 20–60% lower during backward gait compared to the literature ([45,46] measured on force plates). For step width, force plate values during backward gait were in line with the literature [45], whereas the mocap system values were lower (~29%), which is likely related to the calibration, as mentioned in Section 4.1. For tandem gait, Kronenbuerger et al. [47] reported lower cadence values in tandem gait compared to our study (~34%, see Table 9), but they used a different study setting with predetermined gait speed. Rao et al. [43] used a force plate in healthy older adults (mean age 84 years) and also found slightly lower values for cadence, velocity and step width in the tandem gait compared to our values, likely related to the age difference between both cohorts. Importantly, in a tandem gait, the heel of one foot is normally placed directly in front of the toes of the other foot. In our study, a hand's width of space had to be left between the feet to allow the force plate to distinguish between both feet. This difference may explain the higher cadence and velocity values found in our study.

There were significant improvements for some of the variables between the pre- and post-training study visits. For normal gait, the force plate analysis revealed improvement in all gait variables after training, whereas the mocap system only revealed an improvement in two variables after training (stride time, cadence) and the smartphone did not show a significant improvement. For backward gait, an improvement was shown for the velocity variable of both force plate and mocap systems. For tandem gait, an improvement after training was found for the two variables of stride time and cadence in the mocap system only.

In the best case, all systems would have shown significant changes over time in the same variables. However, the differences between the systems may result from (a) reduced statistical power due to a lower number of valid values included in the statistical analysis (as for the smartphone data), and (b) higher variability observed for smartphone data; both of which affect the outcome of the statistical tests. Regarding the two standard systems, the force plate detected more changes in normal gait over time in healthy adult subjects undergoing a training period of three weeks. By comparison, only the mocap system detected changes in tandem gait. One reason for these differences could be that the hardware and software used for the force plate are more accurate for normal walking (because it uses position data, see Section 4.1), but had difficulties distinguishing right and left feet in the tandem gait, whereas the manual detection of steps in the tandem gait was more controllable in the mocap system analysis. Nevertheless, a general improvement in gait variables was observed across all gait analysis systems.

The observed improvements were expected and desirable changes in terms of improved gait performance after a training intervention, and have also been described in several patient studies with various disorders such as PD [15,56] and stroke [57], or for healthy (mostly older) adults after different kinds of training [58–62].

Of note, the observed improvement between pre- and post-training visits is most probably caused by the training performed between these visits. However, a control group undergoing the measurements at T1 and T2 without any training in the interim was missing and, therefore, a learning effect cannot be entirely excluded. To confirm and substantiate the positive effects of this study, further investigation, including a control group, would be reasonable in future.

4.4. Balance Performance

For normal stance, mean values of balance performance (ellipse area) measured with a force plate were comparable with corresponding values of healthy adults in the literature [20,63]. Although, for narrow stance, the velocity values of our study were also comparable or slightly higher than the values of the studies cited above, the values for the ellipse area differed. This is most likely due to methodological differences regarding the calculation of this variable, which is not specified in the studies mentioned above. Pomarino et al. [63] mentioned, however, that their balance measures were averaged over the recording time. In our study, averaged ellipse area values for normal stance were 24 mm^2, 50.7 mm^2 and 3.3 mm^3 (force plate, mocap system and smartphone, respectively), which again is comparable to or slightly lower than in the studies by Nusseck and Spahn [20] and Pomarino et al. [63], who measured with force plates.

For the other stance tasks, reference values for healthy adults in the literature are scarce. One study reported an ellipse area of 138 mm^2 for the single leg stance in a control group of older adults [64], whereas we found values of 878 mm^2, 3860 mm^2 and 384 mm^3 in our study (averaged values per second: 29 mm^2, 129 mm^2 and 13 mm^3). However, it is unclear if the values were indeed averaged in the cited study. If so, the values in our study were lower compared to those in the literature, possibly due to a lower mean age of the participants. Values for the velocity balance variable were only reported separately for mediolateral and anteroposterior directions [64] and are thus not comparable to our values. Terra et al. [38] examined the same stance tasks we used in PD patients, using a force plate, and described an increase in the values for the COM ellipse area and velocity with the level of difficulty of the respective stance tasks, ranging from narrow stance to narrow stance with eyes closed, followed by tandem stance and, finally, single leg stance. This is consistent with our results regarding the velocity variable obtained with the force plate, whereas, for the other systems, the order of the stance tasks varied (see Figure 5).

Regarding the training effects, the statistical analysis did not show a significant improvement in balance performance between pre- and post-training measurements from T1 to T2 (see Figure 5). In contrast to the gait tasks, where small improvements in performance were observed for all variables (even though not always reaching statistical significance), the pattern of observed changes in stance tasks was more heterogeneous (see Table A1, Appendix A). In contrast, an improvement was reported in the literature for different patient groups, e.g., for PD patients [65] or for children with cerebral palsy [21,27] and healthy older adults [66], and for younger adults [67] after a training intervention. Cadore et al. [68] also summarized in their review that most balance trainings in older adults with physical frailty led to enhancements in balance. However, methods, outcome measures and training interventions were highly heterogeneous among the cited studies, impeding their comparability.

4.5. Summary

Agreement between the three gait analysis systems was higher for gait variables than for balance variables. With the exception of the step width variable, both standard methods showed an excellent agreement between the values of the analyzed gait variables, especially for the normal gait task, followed by tandem and backward gait tasks. In particular, for the stride time and cadence variables, values obtained with the smartphone showed a strong correlation with values obtained with both standard systems, whereas correlations for the gait velocity variable were considerably weaker, especially for tandem and backward gait. Improvements (by percentage change) were consistently visible across all gait tasks and all three applied gait analysis systems. However, significant changes over time were only found for gait variables obtained from the force plate and mocap systems. In contrast, changes in balance variables over time yielded a highly heterogeneous pattern without clear improvement across stance tasks and applied systems. Furthermore, participant-reported outcome measures did not reveal any changes over time, which may be due to the already high level of “general well-being” at the study onset.

According to the results of our research, there is a high level of agreement between the devices used in the laboratory and smartphones. This finding is consistent with the findings of earlier studies [69,70]. The fact that smartphones and smartwatches can be put to use in everyday settings is the primary advantage of using such devices. Because of this capability, patients can be monitored in (near) real time and over extended time periods such as months and years. In addition, the vast number of people who own smartphones makes it possible to use these devices as an excellent source for crowdsourcing, regardless of the physical location of the users. However, there are additional considerations such as misunderstanding and following of instructions, effect of motivation, learning effects and misplacement or orientation of devices for at-home usage settings and self-administered protocols, both of which have the potential to affect the validity and reliability of the data collected [71].

Since improvements were found only for gait performance, the applicability of smartphones as a measurement system seems to be particularly useful in disorders in which the gait is impaired, such as PD and ataxia [13,14]. Stride time and cadence measured with the smartphone were found to have a high agreement with the measurements of the standard analysis systems and are variables that differentiate patients from healthy controls [13] or that might improve after an intervention [15]. For this reason, they seem to be eligible variables for future smartphone studies in home-based environments. Future studies should investigate the most effective intervention program and should combine a longer time frame for exercise interventions with major efforts to maintain or even improve study compliance.

5. Conclusions

Our analysis showed that measuring gait and balance performance in healthy adults with wearable devices, such as smartphones, produced comparable results for the stride time and cadence variables compared to measurements with standard gait analysis systems such as the force plate or mocap systems, whereas results for gait velocity were less convincing. Potentially, adjustments may have to be made in the data evaluation for the calculation of velocity to achieve better agreement.

Although the positive influence of three weeks of gait and balance training on gait performance in healthy adults was noteworthy, comparable improvements were found for all three gait analysis systems in gait parameters. However, only the force plate and the mocap systems were able to detect significant changes over time during the gait tasks. In contrast to the motor performance, no improvement was found for the questionnaire scores. To ensure that the improvement is indeed the effect of the training and not a test–retest effect, a further study including a control group which does not take part in a training intervention is required.

Reference values for gait and balance variables in healthy adults are currently scarce in the literature. For future analyses, the number of comparable gait and balance variables can be increased to obtain a more detailed overview of reference values of healthy adults and to compare these values with patient data (e.g., patients with movement disorders). Ellis et al. [13] also suggested that many more consecutive steps (e.g., more than 100 steps) are required to reliably detect differences in gait performance. This is not possible when using force plates with a limited length, but seems to be an interesting set-up for further smartphone-based analyses.

Author Contributions: Conceptualization, C.R. and M.M.; methodology, C.R. and M.M.; software, C.R., M.B. and J.D.; validation, M.S.F., M.B., J.D. and M.M.; formal analysis, C.R., M.S.F. and J.D.; investigation, C.R.; resources, M.B., K.A., A.S., J.D. and M.M.; data curation, C.R. and M.B.; writing—original draft preparation, C.R., M.B., J.D. and M.M.; writing—review and editing, M.S.F., A.S. and K.A.; visualization, C.R. and J.D.; supervision, M.M.; project administration, C.R. and M.M.; funding acquisition, K.A. All authors have read and agreed to the published version of the manuscript.

Funding: Open access publication was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—491111487. This research received no additional funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Heinrich-Heine University in Düsseldorf (Germany) in April 2021.

Informed Consent Statement: Written informed consent has been obtained from the patient(s) to publish this paper.

Data Availability Statement: The sensor data presented in this study are openly available in the pedestrian dynamics data archive "Clara Rentz, Mehran Sahandi Far, Maik Boltes, Alfons Schnitzler, Katrin Amunts, Juergen Dukart, Martina Minnerop. System comparison for gait and balance monitoring used for the evaluation of a home-based training. Forschungszentrum Jülich, 2022" at https://doi.org/10.34735/ped.2021.1, accessed on 30 June 2022.

Acknowledgments: We would especially like to thank all the participants who took part in this study. We also thank PhysioStützpunkt (Köln, Germany) for the creation of gait and balance training videos.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Gait Performance

In Table A1, values of all gait variables are displayed before training (T1) and after training (T2) for all three systems. Significant differences in time were found for normal gait (force plate). In detail, significant differences within the post-hoc test were found for all variables within normal gait and two variables within backward gait (force plate); two variables within normal gait, one within backward gait and three within tandem gait (mocap system); and two variables within normal gait and one within tandem gait (smartphone).

Table A1. Differences in mean between the first (T1) and second study visit (T2) for the gait variables of all three gait analysis systems. The percentage change is indicated in "Δ %". Bold font indicates a significant difference in time (T1-T2, $p < 0.013$ for the force plate and mocap systems, $p < 0.017$ for smartphone) and bold plus italic font indicates a difference in time in the Wilcoxon rank test ($p < 0.013/p < 0.017$). Italic font indicates the implementation of a Wilcoxon rank test. An asterisk marks all significant values in general. Min. = minimum, max. = maximum, SD = standard deviation.

			T1			**T2**			***p***	**Δ %**
			n	**Mean± SD**	**Range**	***n***	**Mean ± SD**	**Range**		
Force plate	NG	stride time [s]	25	**1.20 ± 0.13**	0.97–1.55	20	**1.13 ± 0.10**	0.91–1.29	**0.003** *	−6.15
		Cadence [steps/s]	25	**1.70 ± 0.17**	1.30–2.08	20	**1.80 ± 0.18**	1.55–2.20	**0.002** *	5.89
		Velocity [m/s]	25	**0.98 ± 0.14**	0.64–1.28	20	**1.09 ± 0.12**	0.92–1.42	**0.002** *	11.01
		step width [cm]	25	***11.64 ± 2.60***	7–16	20	***10.65 ± 2.50***	7–15	***0.004*** *	−8.51
	BG	stride time [s]	25	1.22 ± 0.13	1.04–1.56	20	1.17 ± 0.12	0.94–1.37	0.027	−4.01
		Cadence [steps/s]	25	1.66 ± 0.16	1.32–1.92	20	1.73 ± 0.18	1.47–2.12	0.028	4.24
		Velocity [m/s]	25	**0.69 ± 0.09**	0.53–0.86	20	**0.76 ± 0.09**	0.61–0.92	**0.005** *	9.43
		step width [cm]	25	18.08 ± 3.19	10–24	20	17.45 ± 3.20	12–24	0.203	−3.48
	TG	stride time [s]	20	*1.66 ± 0.31*	1.19–2.44	19	*1.61 ± 0.35*	1.00–2.44	*0.031*	−2.93
		Cadence [steps/s]	21	1.23 ± 0.24	0.68–1.68	19	1.33 ± 0.26	0.85–2.02	0.019	8.59
		Velocity [m/s]	21	*0.45 ± 0.12*	0.22–0.72	18	*0.49 ± 0.13*	0.25–0.83	*0.027*	7.81
		step width [cm]	21	*2.24 ± 1.04*	1–5	19	*2.00 ± 0.94*	1–4	*0.624*	−10.71
Mocap system	NG	stride time [s]	24	**1.18 ± 0.13**	0.94–1.51	21	**1.11 ± 0.10**	0.93–1.28	**0.002** *	−6.36
		Cadence [steps/s]	24	**1.71 ± 0.18**	1.32–2.13	21	**1.82 ± 0.17**	1.56–2.14	**0.001** *	6.42
		Velocity [m/s]	24	0.97 ± 0.15	0.59–1.27	21	1.03 ± 0.17	0.65–1.39	0.071	6.48
		step width [cm]	24	10.60 ± 3.42	5.30–15.99	21	9.27 ± 3.48	1.88–16.83	0.266	−12.5

Table A1. *Cont.*

			T1			**T2**			*p*	**Δ %**
			n	**Mean± SD**	**Range**	***n***	**Mean ± SD**	**Range**		
	BG	stride time [s]	24	1.21 ± 0.11	1.03–1.46	21	1.16 ± 0.11	0.94–1.35	0.073	−4.45
		Cadence [steps/s]	24	1.66 ± 0.15	1.37–1.95	21	1.74 ± 0.18	1.48–2.14	0.074	4.79
		Velocity [m/s]	24	**0.66 ± 0.12**	0.31–0.84	21	**0.75 ± 0.10**	0.58–0.89	**0.007 ***	13.91
		step width [cm]	24	11.86 ± 3.40	6.24–19.67	21	11.53 ± 3.70	2.45–17.88	0.676	−2.79
	TG	stride time [s]	24	**1.76 ± 0.42**	1.17–3.11	21	**1.49 ± 0.23**	1.00–1.96	**0.003 ***	−15.33
		Cadence [steps/s]	24	**1.19 ± 0.25**	0.64–1.70	20	**1.35 ± 0.18**	1.02–1.69	**0.001 ***	12.72
		Velocity [m/s]	24	*0.40 ± 0.17*	0.15–0.98	20	*0.44 ± 0.13*	0.19–0.80	*0.024*	10.28
		step width [cm]	22	*2.44 ± 1.06*	0.72–5.67	21	*2.84 ± 1.73*	0.81–7.48	*0.601*	16.1
Smartphone	NG	stride time [s]	23	1.20 ± 0.12	1.00–1.46	16	1.14 ± 0.10	0.94–1.31	0.019	−5.26
		Cadence [steps/s]	23	1.67 ± 0.18	1.32–2.09	16	1.76 ± 0.16	1.52–2.08	0.019	5.39
		Velocity [m/s]	23	1.18 ± 0.51	0.03–2.11	16	1.33 ± 0.39	0.73–2.11	0.639	12.71
	BG	stride time [s]	23	1.23 ± 0.09	1.08–1.42	15	1.23 ± 0.13	1.05–1.43	0.93	0
		Cadence [steps/s]	23	1.62 ± 0.11	1.40–1.84	15	1.62 ± 0.17	1.33–1.89	0.884	0
		Velocity [m/s]	23	*0.55 ± 0.43*	0.07–1.26	15	*0.62 ± 0.44*	0.07–1.40	*0.084*	12.73
	TG	stride time [s]	19	1.67 ± 0.27	1.40–2.47	15	1.51 ± 0.20	1.21–1.99	0.065	−10.6
		Cadence [steps/s]	19	1.23 ± 0.17	0.82–1.43	15	1.35 ± 0.18	1.00–1.65	0.048	9.76
		Velocity [m/s]	19	*0.20 ± 0.22*	0.01–0.66	15	*0.31 ± 0.25*	0.06–0.82	*0.333*	55

* Correlation is significant, *p*-levels vary.

Appendix A.2. Balance Performance

Significant differences between balance variables measured at the first and second study visit were less frequent than those between gait variables. Significant differences in time were found only for tandem stance (force plate). Significant differences in the post-hoc test were present for the COM velocity in the tandem stance (force plate).

Table A2. Differences in mean between the first (T1) and second study visit (T2) for the balance variables of all three gait analysis systems. Bold font indicates a significant difference in time (T1-T2, $p < 0.05$) and italic font indicates a difference in time in the post-hoc test only ($p < 0.05$). An asterisk marks all significant values in general. COM = center of mass, min. = minimum, max. = maximum, NS = narrow stance, NSEc = narrow stance with eyes closed, SD = standard deviation, SS = single leg stance, TS = tandem stance.

			T1			**T2**			***p***	**Δ %**
			n	**Mean ± SD**	**Range**	***n***	**Mean ± SD**	**Range**		
Force plate	NS	COM ellipse [mm^2]	25	*719.92 ± 307.54*	206.0–1439.0	20	*688.30 ± 352.60*	256.0–1826.0	*0.121*	−4.39
		COM velocity [mm/s]	25	15.60 ± 4.02	9.0–23.0	20	16.30 ± 5.30	8.0–31.0	0.744	4.49
	TS	COM ellipse [mm^2]	25	*1430.44 ± 853.08*	336.0–3348.0	19	*1075.21 ± 594.77*	227.0–2314.0	*0.277*	−24.83
		COM velocity [mm/s]	24	*52.33 ± 17.93*	28.0–107.0	20	*50.15 ± 29.78*	22.0–135.0	*0.025*	−4.17
	NSEc	COM ellipse [mm^2]	24	*981.33 ± 366.76*	296.0–1622.0	20	*960.10 ± 400.45*	345.0–1730.0	*0.526*	−2.16
		COM velocity [mm/s]	25	27.64 ± 7.48	11.0–42.0	20	25.60 ± 8.52	12.0–48.0	0.094	−7.38
	SS	COM ellipse [mm^2]	20	*878.05 ± 221.37*	439.0–1255.0	20	*977.80 ± 447.48*	394.0–2345.0	*0.601*	11.36
		COM velocity [mm/s]	24	*53.63 ± 26.48*	24.0–111.0	20	*47.85 ± 21.84*	22.0–109.0	*0.082*	−10.78
Mocap system	NS	COM ellipse [mm^2]	24	*1521.86 ± 772.73*	312.2–3628.9	21	*1358.79 ± 727.22*	522.4–3527.5	*0.145*	−10.72
		COM velocity [mm/s]	24	*6.58 ± 1.53*	4.76–10.48	20	*6.44 ± 1.48*	3.7–10.1	*0.232*	−2.13
	TS	COM ellipse [mm^2]	23	*1515.35 ± 948.28*	263.6–4095.1	20	*1397.48 ± 681.17*	376.2–2609.4	*0.575*	−7.78
		COM velocity [mm/s]	22	*8.55 ± 1.81*	5.1–11.6	21	*9.32 ± 3.77*	4.4–19.3	*0.881*	9.01
Mocap system	NSEc	COM ellipse [mm^2]	23	*1730.55 ± 655.97*	754.3–3138.0	21	*1542.95 ± 829.00*	528.7–3467.2	*0.167*	−10.84
		COM velocity [mm/s]	24	8.72 ± 2.41	5.47–16.37	21	7.76 ± 2.03	3.9–11.4	0.075	−11.01
	SS	COM ellipse [mm^2]	20	*3859.69 ± 3862.79*	466.0–15,835.0	18	*2710.43 ± 2320.89*	434.8–10,074.4	*1*	−29.78
		COM velocity [mm/s]	21	*13.07 ± 6.45*	6.6–28.9	20	*11.85 ± 4.17*	6.4–21.9	*0.557*	−9.33
Smartphone	NS	COM ellipse [mm^2]	16	*97.74 ± 119.94*	0.2–415.9	11	*785.18 ± 1224.51*	0.0–3510.4	*0.753*	703.34
		COM velocity [mm/s]	21	*48.20 ± 27.31*	12.5–112.5	12	*62.96 ± 27.88*	34.1–114.8	*0.333*	30.62
	TS	COM ellipse [mm^2]	18	*967.22 ± 1345.26*	1.0–4402.8	7	*165.23 ± 147.48*	15.9–393.2	*0.043*	−485.38
		COM velocity [mm/s]	20	*59.57 ± 37.42*	18.8–168.8	10	*72.73 ± 46.74*	15.6–178.2	*0.953*	22.09
	NSEc	COM ellipse [mm^2]	16	*49.05 ± 42.87*	0.3–156.0	9	*500.58 ± 715.16*	24.3–1989.0	*0.173*	920.55
		COM velocity [mm/s]	22	*54.67 ± 35.23*	12.1–139.8	9	*49.46 ± 34.13*	22.9–128.5	*0.26*	−10.53
	SS	COM ellipse [mm^2]	16	*383.69 ± 495.16*	15.9–1818.6	9	*591.83 ± 724.82*	10.9–1743.2	*0.31*	54.25
		COM ellipse [mm^2]	16	*97.74 ± 119.94*	0.2–415.9	11	*785.18 ± 1224.51*	0.0–3510.4	*0.753*	703.34

Appendix A.3. Parameter Optimization

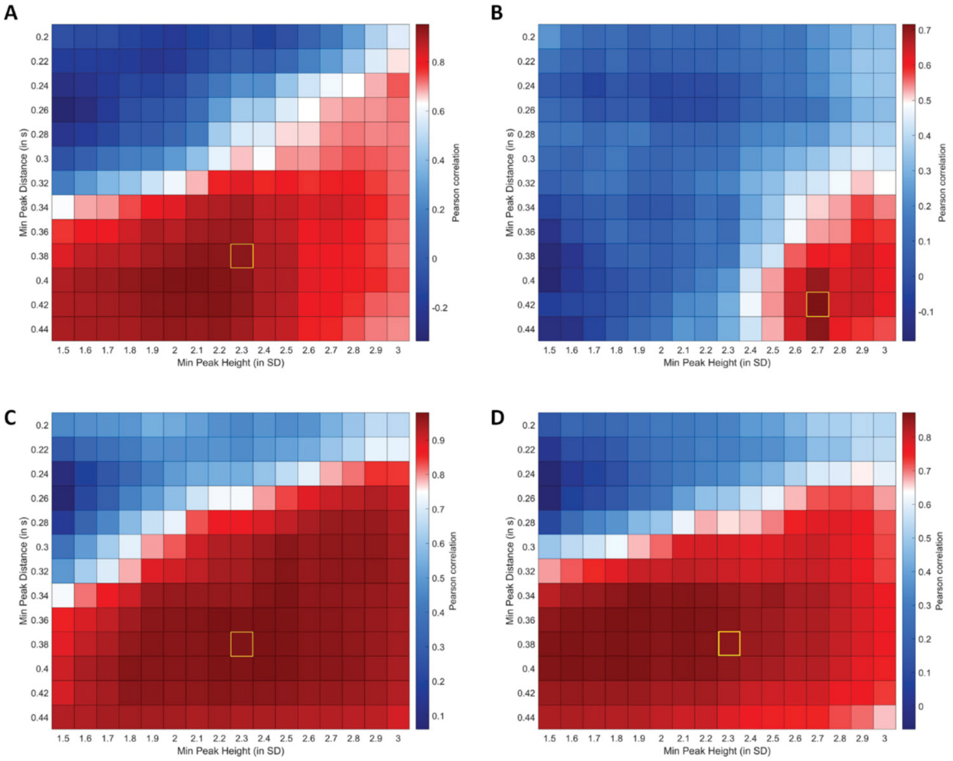


Figure A1. Results of MSI correlation analyses for smartphone step detection parameter optimization. (**A**) Correlation matrix between MSI for backward gait derived from force plate and smartphone data. (**B**) Correlation matrix between MSI for tandem gait derived from force plate and smartphone data. (**C**) Correlation matrix between MSI for normal gait derived using force plate and smartphone data. (**D**) Correlation matrix between MSI for normal gait derived from manual labeling and the automated step detection using smartphone data. Yellow box highlights the final parameters used for subsequent cross-platform comparisons.

Appendix A.4. Ellipsoid Calculation

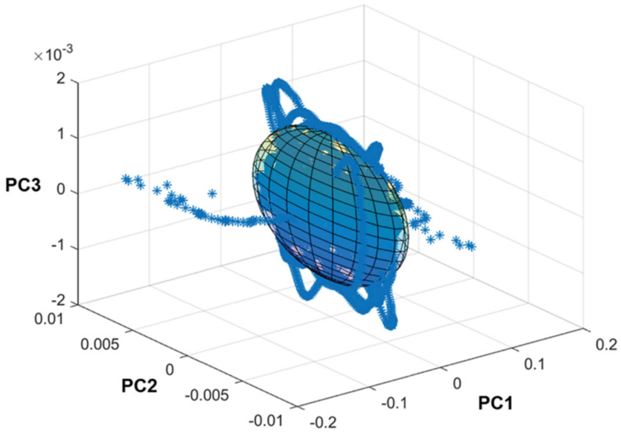


Figure A2. Exemplary visualization of the principal component-based ellipsoid calculation for balance data collected using the JTrack smartphone platform.

References

1. Cruz-Jimenez, M. Normal Changes in Gait and Mobility Problems in the Elderly. *Phys. Med. Rehabil. Clin. N. Am.* **2017**, *28*, 713–725. [CrossRef] [PubMed]
2. Moon, Y.; Sung, J.; An, R.; Hernandez, M.E.; Sosnoff, J.J. Gait variability in people with neurological disorders: A systematic review and meta-analysis. *Hum. Mov. Sci.* **2016**, *47*, 197–208. [CrossRef] [PubMed]
3. Nonnekes, J.; Goselink, R.J.M.; Růžička, E.; Fasano, A.; Nutt, J.G.; Bloem, B.R. Neurological disorders of gait, balance and posture: A sign-based approach. *Nat. Rev. Neurol.* **2018**, *14*, 183–189. [CrossRef] [PubMed]
4. Balestrino, R.; Schapira, A.H.V. Parkinson disease. *Eur. J. Neurol.* **2020**, *27*, 27–42. [CrossRef]
5. Jayadev, S.; Bird, T.D. Hereditary ataxias: Overview. *Genet. Med.* **2013**, *15*, 673–683. [CrossRef]
6. Marsden, J.F. Cerebellar ataxia. In *Handbook of Clinical Neurology: Balance, Gait, and Falls*; Day, B.L., Lord, S.R., Eds.; Elsevier: Amsterdam, The Netherlands, 2018; pp. 261–281, ISBN 0072-9752.
7. Kuang, R.; Ye, Y.; Chen, Z.; He, R.; Savović, I.; Djordjevich, A.; Savović, S.; Ortega, B.; Marques, C.; Li, X.; et al. Low-cost plastic optical fiber integrated with smartphone for human physiological monitoring. *Opt. Fiber Technol.* **2022**, *71*, 102947. [CrossRef]
8. De Farias, F.A.C.; Dagostini, C.M.; Bicca, Y.d.A.; Falavigna, V.F.; Falavigna, A. Remote Patient Monitoring: A Systematic Review. *Telemed. J. E-Health Off. J. Am. Telemed. Assoc.* **2020**, *26*, 576–583. [CrossRef]
9. Thierfelder, A.; Seemann, J.; John, N.; Harmuth, F.; Giese, M.; Schüle, R.; Schöls, L.; Timmann, D.; Synofzik, M.; Ilg, W. Real-Life Turning Movements Capture Subtle Longitudinal and Preataxic Changes in Cerebellar Ataxia. *Mov. Disord.* **2022**, *37*, 1047–1058. [CrossRef]
10. Ilg, W.; Seemann, J.; Giese, M.; Traschütz, A.; Schöls, L.; Timmann, D.; Synofzik, M. Real-life gait assessment in degenerative cerebellar ataxia: Toward ecologically valid biomarkers. *Neurology* **2020**, *95*, e1199–e1210. [CrossRef]
11. Shah, V.V.; Rodriguez-Labrada, R.; Horak, F.B.; McNames, J.; Casey, H.; Hansson Floyd, K.; El-Gohary, M.; Schmahmann, J.D.; Rosenthal, L.S.; Perlman, S.; et al. Gait Variability in Spinocerebellar Ataxia Assessed Using Wearable Inertial Sensors. *Mov. Disord.* **2021**, *36*, 2922–2931. [CrossRef]
12. Petraglia, F.; Scarcella, L.; Pedrazzi, G.; Brancato, L.; Puers, R.; Costantino, C. Inertial sensors versus standard systems in gait analysis: A systematic review and meta-analysis. *Eur. J. Phys. Rehabil. Med.* **2019**, *55*, 265–280. [CrossRef]
13. Ellis, R.J.; Ng, Y.S.; Zhu, S.; Tan, D.M.; Anderson, B.; Schlaug, G.; Wang, Y. A Validated Smartphone-Based Assessment of Gait and Gait Variability in Parkinson's Disease. *PLoS ONE* **2015**, *10*, e0141694. [CrossRef]
14. Schmitz-Hübsch, T.; Brandt, A.U.; Pfueller, C.; Zange, L.; Seidel, A.; Kühn, A.A.; Paul, F.; Minnerop, M.; Doss, S. Accuracy and repeatability of two methods of gait analysis-GaitRite™ und Mobility Lab™-in subjects with cerebellar ataxia. *Gait Posture* **2016**, *48*, 194–201. [CrossRef]
15. Conradsson, D.; Löfgren, N.; Nero, H.; Hagströmer, M.; Ståhle, A.; Lökk, J.; Franzén, E. The Effects of Highly Challenging Balance Training in Elderly with Parkinson's Disease: A Randomized Controlled Trial. *Neurorehabil. Neur. Rep.* **2015**, *29*, 827–836. [CrossRef]

16. Giardini, M.; Nardone, A.; Godi, M.; Guglielmetti, S.; Arcolin, I.; Pisano, F.; Schieppati, M. Instrumental or Physical-Exercise Rehabilitation of Balance Improves Both Balance and Gait in Parkinson's Disease. *Neur. Plastic.* **2018**, *2018*, 5614242. [CrossRef]
17. Podsiadlo, D.; Richardson, S. The timed "Up & Go": A test of basic functional mobility for frail elderly persons. *J. Am. Geriatr. Soc.* **1991**, *39*, 142–148. [CrossRef]
18. Tchelet, K.; Stark-Inbar, A.; Yekutieli, Z. Pilot Study of the EncephaLog Smartphone Application for Gait Analysis. *Sensors* **2019**, *19*, 5179. [CrossRef]
19. Baker, N.; Gough, C.; Gordon, S. Classification of Balance Assessment Technology: A Scoping Review of Systematic Reviews. *Stud. Health Technol. Inform.* **2020**, *268*, 45–59. [CrossRef]
20. Nusseck, M.; Spahn, C. Comparison of Postural Stability and Balance Between Musicians and Non-musicians. *Front. Psychol.* **2020**, *11*, 1253. [CrossRef]
21. Wan, G.; Hsieh, H.-C.; Lin, C.-H.; Lin, H.-Y.; Lin, C.-Y.; Chiu, W.-H. An Accessible Training Device for Children with Cerebral Palsy. *IEEE Trans. Neur. Syst. Rehabil. Eng.* **2021**, *29*, 1252–1258. [CrossRef]
22. Morenilla, L.; Márquez, G.; Sánchez, J.A.; Bello, O.; López-Alonso, V.; Fernández-Lago, H.; Fernández-del-Olmo, M.Á. Postural Stability and Cognitive Performance of Subjects With Parkinson's Disease During a Dual-Task in an Upright Stance. *Front. Psychol.* **2020**, *11*, 1256. [CrossRef]
23. Sun, R.; Moon, Y.; McGinnis, R.S.; Seagers, K.; Motl, R.W.; Sheth, N.; Wright, J.A.; Ghaffari, R.; Patel, S.; Sosnoff, J.J. Assessment of Postural Sway in Individuals with Multiple Sclerosis Using a Novel Wearable Inertial Sensor. *Digit. Biomark.* **2018**, *2*, 485958. [CrossRef]
24. Ajzenman, H.F.; Standeven, J.W.; Shurtleff, T.L. Effect of hippotherapy on motor control, adaptive behaviors, and participation in children with autism spectrum disorder: A pilot study. *Am. J. Occup. Ther.* **2013**, *67*, 653–663. [CrossRef]
25. Cyma-Wejchenig, M.; Tarnas, J.; Marciniak, K.; Stemplewski, R. The Influence of Proprioceptive Training with the Use of Virtual Reality on Postural Stability of Workers Working at Height. *Sensors* **2020**, *20*, 3731. [CrossRef]
26. In, T.; Lee, K.; Song, C. Virtual Reality Reflection Therapy Improves Balance and Gait in Patients with Chronic Stroke: Randomized Controlled Trials. *Med. Sci. Monit.* **2016**, *22*, 4046–4053. [CrossRef]
27. Hsieh, H.-C. Preliminary Study of the Effect of Training With a Gaming Balance Board on Balance Control in Children With Cerebral Palsy: A Randomized Controlled Trial. *Am. J. Phys. Med. Rehabil.* **2020**, *99*, 142–148. [CrossRef]
28. Far, M.S.; Stolz, M.; Fischer, J.M.; Eickhoff, S.B.; Dukart, J. JTrack: A Digital Biomarker Platform for Remote Monitoring in Neurological and Psychiatric Diseases. *arXiv* **2021**. [CrossRef]
29. Xsens Technologies, B.V. MVN User Manual: Document MVNManual, Revision Z. 2020. Available online: https://www.xsens.com (accessed on 9 June 2022).
30. Schepers, M.; Giuberti, M.; Bellusci, G. Xsens MVN: Consistent Tracking of Human Motion Using Inertial Sensing. *Xsens Technol.* **2018**, *1*, 1–8.
31. Kroenke, K.; Spitzer, R.L.; Williams, J.B. The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Int. Med.* **2001**, *16*, 606–613. [CrossRef]
32. Löwe, B.; Spitzer, R.L.; Zipfel, S.; Herzog, W. *PHQ-D. Gesundheitsfragebogen für Patienten: Komplettversion und Kurzform. Autorisierte Deutsche Version des "Prime MD Patient Health Questionnaire (PHQ)"*, 2nd ed.; Pfizer: New York, NY, USA, 2002.
33. Zigmond, A.S.; Snaith, R.P. The hospital anxiety and depression scale. *Acta Psychiatr. Scand.* **1983**, *67*, 361–370. [CrossRef]
34. Hermann-Lingen, C.; Buss, U.; Snaith, R.P. *Hospital Anxiety and Depression Scale–German Version (HADS-D)*; Hans Huber: Bern, Switzerland, 2011.
35. Wydra, G. *Der Fragebogen Zum Allgemeinen Habituellen Wohlbefinden (Fahw Und Fahw-12): Entwicklung Und Evaluation Eines Mehrdimensionalen Fragebogens (5. Überarbeitete Und Erweiterte Version)*; Universität des Saarlandes: Saarbrücken, Germany, 2014.
36. Scholler, G.; Fliege, H.; Klapp, B.F. *SWOP-K9-Fragebogen zu Selbstwirksamkeit-Optimismus-Pessimismus Kurzform*; ZPID (Leibniz Institute for Psychology Information)–Testarchiv: Trier, Germany, 1999.
37. Schott, N. Deutsche Adaptation der "Activities-Specific Balance Confidence (ABC) Scale" zur Erfassung der sturzassoziierten Selbstwirksamkeit. *Zeitschrift für Gerontologie Geriatrie* **2008**, *41*, 475–485. [CrossRef] [PubMed]
38. Terra, M.B.; Da Silva, R.A.; Bueno, M.E.B.; Ferraz, H.B.; Smaili, S.M. Center of pressure-based balance evaluation in individuals with Parkinson's disease: A reliability study. *Physiother. Theory Pract.* **2020**, *36*, 826–833. [CrossRef] [PubMed]
39. Kouvelioti, V.; Kellis, E.; Kofotolis, N.; Amiridis, I. Reliability of Single-leg and Double-leg Balance Tests in Subjects with Anterior Cruciate Ligament Reconstruction and Controls. *Res. Sports Med.* **2015**, *23*, 151–166. [CrossRef] [PubMed]
40. Dancey, C.P.; Reidy, J. *Statistics without Maths for Psychology*, 4th ed.; Pearson Prentice Hall: Harlow, UK, 2008; ISBN 9780132051606.
41. Pawik, Ł.; Fink-Lwow, F.; Pajchert Kozłowska, A.; Szelerski, Ł.; Żarek, S.; Górski, R.; Pawik, M.; Urbanski, W.; Reichert, P.; Morasiewicz, P. Assessment of Gait after Treatment of Tibial Nonunion with the Ilizarov Method. *Int. J. Environ. Res. Public Health* **2021**, *18*, 4217. [CrossRef]
42. Kasović, M.; Štefan, L.; Borovec, K.; Zvonař, M.; Cacek, J. Effects of Carrying Police Equipment on Spatiotemporal and Kinetic Gait Parameters in First Year Police Officers. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5750. [CrossRef]
43. Rao, A.K.; Gillman, A.; Louis, E.D. Quantitative gait analysis in essential tremor reveals impairments that are maintained into advanced age. *Gait Posture* **2011**, *34*, 65–70. [CrossRef]
44. Whittle, M.W. *Gait Analysis: An Introduction*, 4th ed.; Butterworth-Heinemann: Oxford, UK, 2007; ISBN 0750688831.

45. Gimunová, M.; Bozděch, M.; Skotáková, A.; Grün, V.; Válková, H. Comparison of forward and backward gait in males with and without intellectual disabilities. *J. Intellect. Disab. Res.* **2021**, *65*, 922–929. [CrossRef]
46. Edwards, E.M.; Kegelmeyer, D.A.; Kloos, A.D.; Nitta, M.; Raza, D.; Nichols-Larsen, D.S.; Fritz, N.E. Backward Walking and Dual-Task Assessment Improve Identification of Gait Impairments and Fall Risk in Individuals with MS. *Mult. Scler. Int.* **2020**, *2020*, 6707414. [CrossRef]
47. Kronenbuerger, M.; Konczak, J.; Ziegler, W.; Buderath, P.; Frank, B.; Coenen, V.A.; Kiening, K.; Reinacher, P.; Noth, J.; Timmann, D. Balance and motor speech impairment in essential tremor. *Cerebellum* **2009**, *8*, 389–398. [CrossRef]
48. Steins, D.; Sheret, I.; Dawes, H.; Esser, P.; Collett, J. A smart device inertial-sensing method for gait analysis. *J. Biomech.* **2014**, *47*, 3780–3785. [CrossRef]
49. Yamada, M.; Aoyama, T.; Mori, S.; Nishiguchi, S.; Okamoto, K.; Ito, T.; Muto, S.; Ishihara, T.; Yoshitomi, H.; Ito, H. Objective assessment of abnormal gait in patients with rheumatoid arthritis using a smartphone. *Rheumatol. Int.* **2012**, *32*, 3869–3874. [CrossRef]
50. Fischetti, F.; Greco, G.; Cataldi, S.; Minoia, C.; Loseto, G.; Guarini, A. Effects of Physical Exercise Intervention on Psychological and Physical Fitness in Lymphoma Patients. *Medicina* **2019**, *55*, 379. [CrossRef]
51. Mikkelsen, K.; Stojanovska, L.; Polenakovic, M.; Bosevski, M.; Apostolopoulos, V. Exercise and mental health. *Maturitas* **2017**, *106*, 48–56. [CrossRef]
52. Chekroud, S.R.; Gueorguieva, R.; Zheutlin, A.B.; Paulus, M.; Krumholz, H.M.; Krystal, J.H.; Chekroud, A.M. Association between physical exercise and mental health in 1·2 million individuals in the USA between 2011 and 2015: A cross-sectional study. *Lancet Psychiatr.* **2018**, *5*, 739–746. [CrossRef]
53. Sherrington, C.; Whitney, J.C.; Lord, S.R.; Herbert, R.D.; Cumming, R.G.; Close, J.C.T. Effective exercise for the prevention of falls: A systematic review and meta-analysis. *J. Am. Geriatr. Soc.* **2008**, *56*, 2234–2243. [CrossRef]
54. Kim, T.W.B.; Gay, N.; Khemka, A.; Garino, J. Internet-Based Exercise Therapy Using Algorithms for Conservative Treatment of Anterior Knee Pain: A Pragmatic Randomized Controlled Trial. *JMIR Rehabil. Assist. Technol.* **2016**, *3*, e12. [CrossRef]
55. Haines, T.P.; Russell, T.; Brauer, S.G.; Erwin, S.; Lane, P.; Urry, S.; Jasiewicz, J.; Condie, P. Effectiveness of a video-based exercise programme to reduce falls and improve health-related quality of life among older adults discharged from hospital: A pilot randomized controlled trial. *Clin. Rehabil.* **2009**, *23*, 973–985. [CrossRef]
56. Atterbury, E.M.; Welman, K.E. Balance training in individuals with Parkinson's disease: Therapist-supervised vs. home-based exercise programme. *Gait Posture* **2017**, *55*, 138–144. [CrossRef]
57. Sungkarat, S.; Fisher, B.E.; Kovindha, A. Efficacy of an insole shoe wedge and augmented pressure sensor for gait training in individuals with stroke: A randomized controlled trial. *Clin. Rehabil.* **2011**, *25*, 360–369. [CrossRef]
58. Batson, C.D.; Brady, R.A.; Peters, B.T.; Ploutz-Snyder, R.J.; Mulavara, A.P.; Cohen, H.S.; Bloomberg, J.J. Gait training improves performance in healthy adults exposed to novel sensory discordant conditions. *Exp. Brain. Res.* **2011**, *209*, 515–524. [CrossRef]
59. Falbo, S.; Condello, G.; Capranica, L.; Forte, R.; Pesce, C. Effects of Physical-Cognitive Dual Task Training on Executive Function and Gait Performance in Older Adults: A Randomized Controlled Trial. *BioMed Res. Int.* **2016**, *2016*, 5812092. [CrossRef] [PubMed]
60. Uematsu, A.; Hortobágyi, T.; Tsuchiya, K.; Kadono, N.; Kobayashi, H.; Ogawa, T.; Suzuki, S. Lower extremity power training improves healthy old adults' gait biomechanics. *Gait Posture* **2018**, *62*, 303–310. [CrossRef] [PubMed]
61. Tasvuran Horata, E.; Cetin, S.Y.; Erel, S. Effects of individual progressive single- and dual-task training on gait and cognition among older healthy adults: A randomized-controlled comparison study. *Eur. Geriatr. Med.* **2021**, *12*, 363–370. [CrossRef] [PubMed]
62. Nagashima, Y.; Ito, D.; Ogura, R.; Tominaga, T.; Ono, Y. Development of Virtual Reality-based Gait Training System Simulating Personal Home Environment. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 1–5 November 2021.
63. Pomarino, D.; Nawrath, A.; Beyer, J. Altersabhängige Messungen zur posturalen Stabilität gesunder Probanden. *OUP* **2013**, *2*, 420–425.
64. Sun, W.; Wang, L.; Zhang, C.; Song, Q.; Gu, H.; Mao, D. Detraining effects of regular Tai Chi exercise on postural control ability in older women: A randomized controlled trial. *J. Exercise Sci. Fitness* **2018**, *16*, 55–61. [CrossRef]
65. Stożek, J.; Rudzińska, M.; Pustułka-Piwnik, U.; Szczudlik, A. The effect of the rehabilitation program on balance, gait, physical performance and trunk rotation in Parkinson's disease. *Aging Clin. Exp. Res.* **2016**, *28*, 1169–1177. [CrossRef]
66. Gordt, K.; Gerhardy, T.; Najafi, B.; Schwenk, M. Effects of Wearable Sensor-Based Balance and Gait Training on Balance, Gait, and Functional Performance in Healthy and Patient Populations: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. *Gerontology* **2018**, *64*, 74–89. [CrossRef]
67. Zech, A.; Hübscher, M.; Vogt, L.; Banzer, W.; Hänsel, F.; Pfeifer, K. Balance training for neuromuscular control and performance enhancement: A systematic review. *J. Athl. Train.* **2010**, *45*, 392–403. [CrossRef]
68. Cadore, E.L.; Rodríguez-Mañas, L.; Sinclair, A.; Izquierdo, M. Effects of different exercise interventions on risk of falls, gait ability, and balance in physically frail older adults: A systematic review. *Rejuven. Res.* **2013**, *16*, 105–114. [CrossRef]
69. Manor, B.; Yu, W.; Zhu, H.; Harrison, R.; Lo, O.-Y.; Lipsitz, L.; Travison, T.; Pascual-Leone, A.; Zhou, J. Smartphone App-Based Assessment of Gait During Normal and Dual-Task Walking: Demonstration of Validity and Reliability. *JMIR Mhealth Uhealth* **2018**, *6*, e36. [CrossRef]

70. Silsupadol, P.; Teja, K.; Lugade, V. Reliability and validity of a smartphone-based assessment of gait parameters across walking speed and smartphone locations: Body, bag, belt, hand, and pocket. *Gait Posture* **2017**, *58*, 516–522. [CrossRef]
71. Far, M.S.; Eickhoff, S.B.; Goni, M.; Dukart, J. Exploring Test-Retest Reliability and Longitudinal Stability of Digital Biomarkers for Parkinson Disease in the m-Power Data Set: Cohort Study. *J. Med. Int. Res.* **2021**, *23*, e26608.

Paper3: Exploring Test-Retest Reliability and Longitudinal Stability of Digital Biomarkers for Parkinson's Disease in the m-Power Data Set: Cohort Study

Mehran Sahandi Far[1,2], Simon B Eickhoff[1,2], Maria Goni[1,2*], Juergen Dukart[1,2]

[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

Corresponding Author:

Juergen Dukart, PhD
Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7) Research Centre Jülich
Wilhelm-Johnen-Strasse
Jülich, 52425
Germany
Phone: 49 1632874330
Fax: 49 2461611880
Email: juergen.dukart@gmail.com

Published in: Journal of Medical Internet Research 2021 Sep; 23(9): e26608.
https://doi.org/10.2196%2F26608

Own contributions

Writing the manuscript, preparing figures, contributing to the design of the experiment, writing analysis code, statistical data analysis, and contributing to the interpretation of results. Total contribution 80%

Original Paper

Exploring Test-Retest Reliability and Longitudinal Stability of Digital Biomarkers for Parkinson Disease in the m-Power Data Set: Cohort Study

Mehran Sahandi Far[1,2], MA; Simon B Eickhoff[1,2], Prof Dr; Maria Goni[1,2*], PhD; Juergen Dukart[1,2*], PhD

[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

[*]these authors contributed equally

Corresponding Author:
Juergen Dukart, PhD
Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7)
Research Centre Jülich
Wilhelm-Johnen-Strasse
Jülich, 52425
Germany
Phone: 49 1632874330
Fax: 49 2461611880
Email: juergen.dukart@gmail.com

Abstract

Background: Digital biomarkers (DB), as captured using sensors embedded in modern smart devices, are a promising technology for home-based sign and symptom monitoring in Parkinson disease (PD).

Objective: Despite extensive application in recent studies, test-retest reliability and longitudinal stability of DB have not been well addressed in this context. We utilized the large-scale m-Power data set 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 patients with PD and healthy controls (HC).

Methods: Intraclass correlation coefficients were computed to estimate the test-retest reliability of features that also differentiate between patients with 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.

Results: Among the features differing between PD and HC, only a 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 a 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.

Conclusions: These results indicate the need for further development of more standardized, sensitive, and reliable DB for application in self-administered remote studies in patients with PD. Motivational, learning, and other confounders may cause variations in performance that need to be considered in DB longitudinal applications.

(J Med Internet Res 2021;23(9):e26608) doi: 10.2196/26608

KEYWORDS

health sciences; medical research; biomarkers; diagnostic markers; neurological disorders; Parkinson disease; mobile phone

Introduction

Parkinson disease (PD) is primarily characterized by motor signs and 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 interrater variability by relying on self-reporting by patients and caregivers or clinicians' 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 (ie, acceleration, gyroscope, GPS, etc) embedded in smart devices for daily use (ie, smartphone, smartwatch). Such sensor data, alongside 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) [3-5]. 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 for assessing agreement between clinical rating scales such as UPDRS and sensor-driven data to quantify disease severity or intervention effects [4,6-9]. 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 (ie, due to motivation, medication, or other aspects) [10-12]. Other factors such as the effect of users' familiarity with technology and the impact of learning on the performance of measured DB in remote and self-administered 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 [13-18]. However, the existing studies typically focused on a single or a 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 the studies that evaluated DB assessments for remote monitoring of PD, only one reported the test-retest reliability [4]. No PD studies systematically evaluated the test-retest reliability and longitudinal sensitivity of DB in a fully unsupervised and self-administered PD longitudinal setting.

Although various factors such as medication, disease severity, learning effects, bias from self-reporting, inconsistent disease severity, motivational impacts, and design protocols in self-administered studies can affect the long-term stability of DB, little attention has been paid to evaluating the reliability and longitudinal stability of DB in loosely controlled self-administered settings in daily life. Here, we aimed to address these open questions by assessing the test-retest reliability and longitudinal stability of gait, balance, speech, and tapping tasks in patients with PD and a control cohort consisting of healthy volunteers (HC) in an unsupervised and self-administered daily life setting using the large-scale m-Power data set [19].

Methods

Study Cohort

To address the open questions on the performance of DB measures in PD when collected in a self-administered setting in daily life, 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-retest reliability of these features as collected using 4 commonly applied PD tasks (gait, balance, voice, and tapping) in daily life using smartphone in a large cohort of self-reported patients with PD and healthy controls, the m-Power study [19-22]. In addition, we evaluated their sensitivity to learning and medication effects.

Enrolment in the m-Power study was open to adult participants who own an iPhone, are living in the United States, and are comfortable enough with English to read the instructions in the app. Participants were asked to download the app and complete a one-time demographic survey during registration. Demographic data include but are 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 UPDRS Section I (nonmotor experience) and Section II (motor experience), as well as the Parkinson's Disease Questionnaire (PDQ-8). All the participants were suggested to complete each task (walking, tapping, voice, and memory) up to 3 times a day for up to 6 months. In addition, self-reported patients with PD were asked to complete the task before medication, after medication, and at another time when they were feeling at their best.

Ethical oversight of the m-Power study was obtained from the Western Institutional Review Board. Prior to signing an electronically rendered traditional informed consent form, prospective participants had to pass a 5-question quiz evaluating their understanding of the study aims, participant rights, and data sharing options. After completing the e-consent process and electronically signing the informed consent form, participants were asked for an email address to which their signed consent form was sent and allowing for verification of their enrolment in the study. Participants were given the option to share their data only with the m-Power study team and partners ("share narrowly") or to share their data more broadly with qualified researchers worldwide, and they had to make an active choice to complete the consent process (no default choice was presented). The data used in our study consist of all individuals who chose to have their data shared broadly.

Data Preprocessing

The m-Power data set is assessed outside of a clinical environment with limited quality control and supervision. All information, including the health history, disease diagnosis, duration, treatment, and survey outcomes, are self-reported. To address these, we excluded participants who did not specify their age, sex, and information on professional diagnosis (if they belong to the PD or HC group) and those with empty, null, or corrupted files. The participants are assigned to the PD or HC group according to their response 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 groups. 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 demographic details are provided in Table 1, and the overall overview of preprocessing steps is displayed in Figure 1A.

Figure 1. Overview of statistical analyses and the preprocessing scheme. (A) Flowchart of preprocessing steps. (B) Flowchart of statistical analyses. (C) Flowchart of number of features at each selection step. HC: healthy controls; ICC: intraclass correlation coefficients; PD: Parkinson disease; rm-ANOVA: repeated-measures analysis of variance.

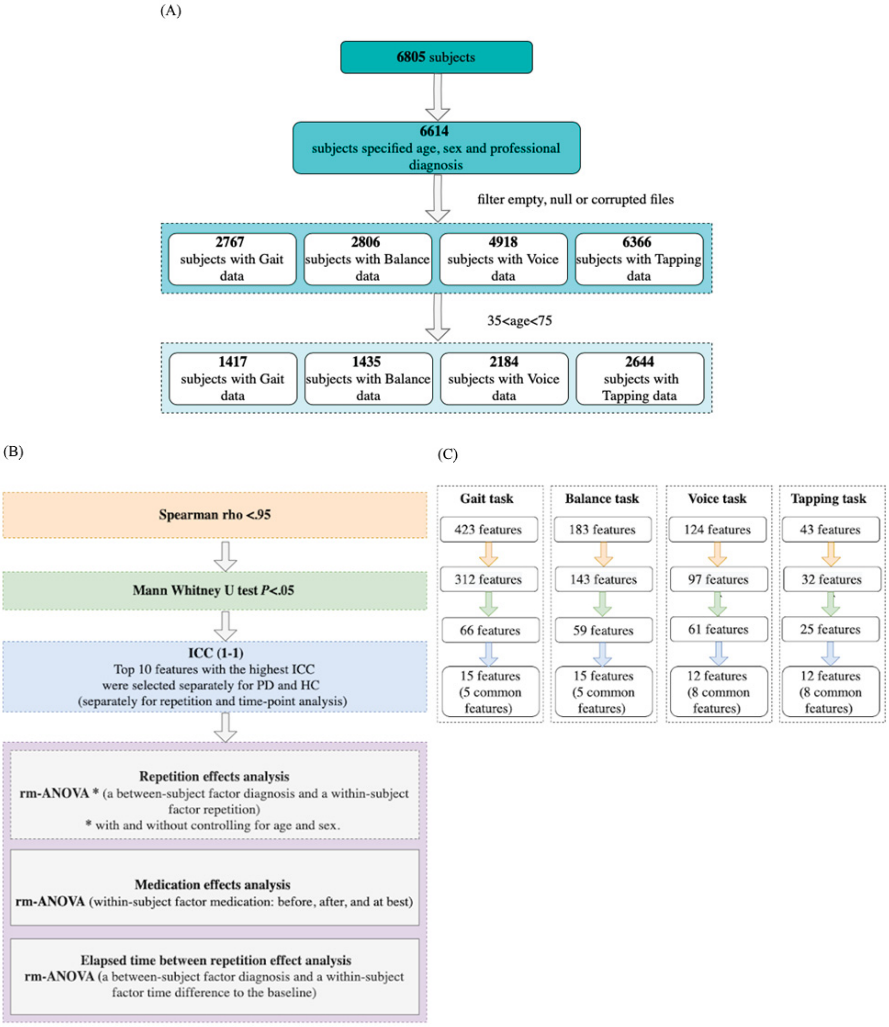

(A)
6805 subjects
6614 subjects specified age, sex and professional diagnosis
filter empty, null or corrupted files
2767 subjects with Gait data
2806 subjects with Balance data
4918 subjects with Voice data
6366 subjects with Tapping data
35<age<75
1417 subjects with Gait data
1435 subjects with Balance data
2184 subjects with Voice data
2644 subjects with Tapping data
(B)
Spearman rho <.95
Mann Whitney U test P<.05
ICC (1-1)
Top 10 features with the highest ICC were selected separately for PD and HC (separately for repetition and time-point analysis)
Repetition effects analysis
rm-ANOVA * (a between-subject factor diagnosis and a within-subject factor repetition)
* with and without controlling for age and sex.
Medication effects analysis
rm-ANOVA (within-subject factor medication: before, after, and at best)
Elapsed time between repetition effect analysis
rm-ANOVA (a between-subject factor diagnosis and a within-subject factor time difference to the baseline)
(C)
Gait task
423 features
312 features
66 features
15 features (5 common features)
Balance task
183 features
143 features
59 features
15 features (5 common features)
Voice task
124 features
97 features
61 features
12 features (8 common features)
Tapping task
43 features
32 features
25 features
12 features (8 common features)

Table 1. Characteristics of study cohorts after data cleaning.

Characteristic	Gait		Balance		Voice		Tapping	
	HC[a]	PD[b]	HC	PD	HC	PD	HC	PD
Sex,[c] n								
Male	655	399	668	401	1042	571	1370	630
Female	152	211	155	211	249	322	304	340
Age (years),[c] mean (SD)	49 (10.60)	60.3 (8.90)	48.9 (10.70)	60.3 (8.90)	47.7 (10.40)	60.1 (9)	46.9 (10.1)	59.9 (9)
UPDRS,[d] mean (SD)	N/A[e]	12.60 (7.11)	N/A	12.53 (7.07)	N/A	12.58 (7.70)	N/A	12.54 (7.73)
UPDRS I, mean (SD)	N/A	4.90 (3.12)	N/A	4.9 (3.11)	N/A	4.93 (3.25)	N/A	4.95 (3.27)
UPDRS II, mean (SD)	N/A	7.76 (5.41)	N/A	7.7 (5.40)	N/A	7.61 (5.70)	N/A	7.56 (5.70)
PDQ-8,[f] mean (SD)	N/A	5.13 (4.72)	N/A	7.07 (4.70)	N/A	5.28 (5.01)	N/A	5.3 (4.96)

[a] HC: healthy controls.

[b] PD: Parkinson disease.

[c] *P*<.001 (two-sample, two-tailed *t* test for age and chi-square test for sex with 95% confidence) for all tasks.

[d] UPDRS: Unified Parkinson's Disease Rating Scale.

[e] N/A: not applicable.

[f] PDQ: Parkinson’s Disease Questionnaire.

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 a total 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 [23]. A detailed explanation of the extracted features, including the respective references, is provided in Tables S1-S4 in Multimedia Appendix 1. For features sharing the same variance (high pairwise correlation: Spearman ρ>0.95), only one of the features was selected randomly for further analyses to reduce the amount of redundant information for each task. Figure 1C summarizes the feature extraction process and the number of features at each selection step.

Gait and Balance

Impairments in gait speed, stride length, and stride time variability are common changes that are linked to PD [24-27]. Instability in postural balance is also considered to be one of the well-reported characteristics associated with PD [15,28-30]. 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-second stay still period. Given a heterogeneity of gait signal lengths across participants, we used a fixed length signal of 10 seconds and selected data from participants who met this criterion, which resulted in 28,150 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 [31,32] (Table S1 in Multimedia Appendix 1). For balance, we used a 15-second time window, trimming the first 5 and the last 10 seconds of the 30-second records to reduce the noise due to the between-task transition period, resulting in 29,050 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 (nontremor) falling in the 0-3.5 Hz band [33] (Table S2 in Multimedia Appendix 1).

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 [16,34,35]. In this task, participants said “aaaah” for about 10 seconds. For voice, 49,676 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 2 seconds of the 10-second interval (Table S3 in Multimedia Appendix 1).

Tapping

Impairment in finger dexterity is another sign associated with PD [36,37]. 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, 55,894 recordings were selected, belonging to 2644 unique participants. Features were computed based on the intertapping distance and interval (Table S4 in Multimedia Appendix 1).

Statistical Analysis

For features to be considered usable for biomarker purposes in longitudinal studies, several criteria are important, including sensitivity to disease signs and symptoms, good test-retest reliability, and robustness against the effects of learning and other longitudinal confounders. To address these criteria, we adopted a stepwise statistical procedure (see Figure 1B for a summary of statistical analyses).

As DB measures are frequently not normally distributed, Mann-Whitney *U* tests were used to identify all features that significantly differ between PD and HC at the first administration (baseline) ($P<.05$). Effect sizes (Cohen *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. We used ICC type 1-1 in our study because individuals were not tested under the same conditions (ie, same device), and reliability was determined from a single measurement. ICC values of 0-0.40 were considered to be poor, 0.40-0.59 to be fair, 0.60-0.74 to be good, and 0.75-1.00 to be excellent [38]. To assess the reliability of each feature, ICC values were computed for different time points versus 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 vs second, third, fourth, and fifth repeat). We then focused our analyses on the top 10 features (as they provide a representative subset of the best performing features) with the highest median ICC values for each group (PD, HC) and tested for their longitudinal stability over time. Results for all features are reported in Multimedia Appendix 1. Features from the PD group are further referred to as "PD features," those from the HC group only as "HC features," and overlapping features from both groups as "common features." We computed repeated-measures analyses of variance (rm-ANOVA) using a mixed factorial design with a between-subject factor diagnosis and a within-subject factor repetition (first, second, third, fourth, and fifth) including their interaction (Equation S1 in Multimedia Appendix 1). Participants who had at least 4 repetitions after baseline (463 for gait, 597 for balance, 1085 for voice, and 1333 for tapping) were included in these analyses. To assess the effects of age and sex on the longitudinal stability of the most reliable features, we repeated all analyses while controlling for age and sex as covariates (Equation S2 in Multimedia Appendix 1). Also, we assessed the impact of elapsed time between repetitions by computing rm-ANOVA using a mixed factorial design with a between-subject factor diagnosis and a within-subject factor elapsed time (calculated as a time difference of each repetition from the baseline in hours) and controlling for age and sex (Equation S3 in Multimedia Appendix 1).

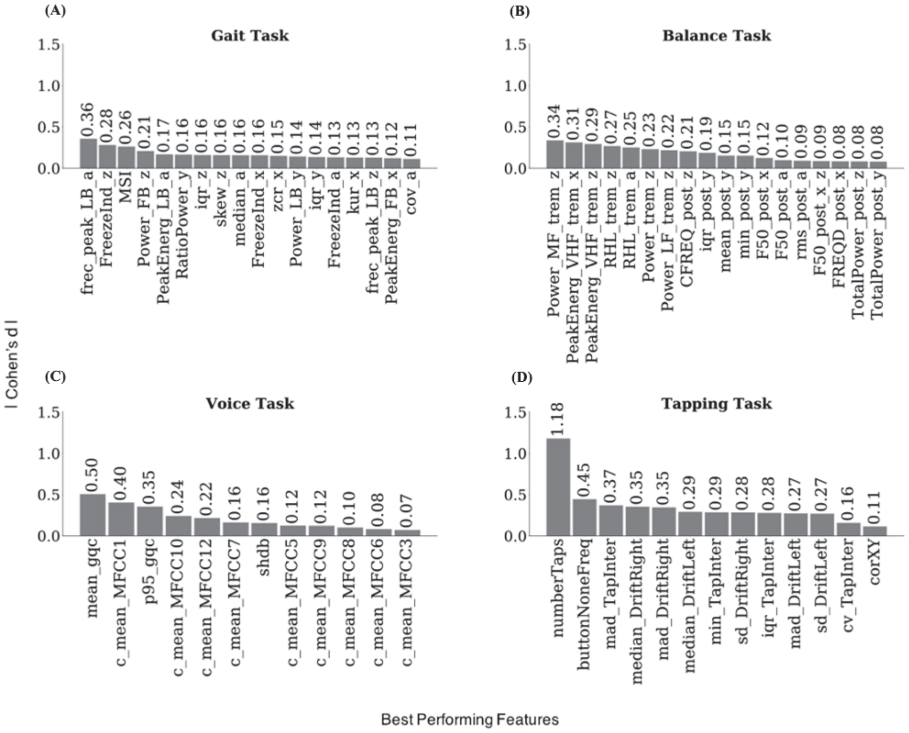
Lastly, we assessed the impact of PD medication by computing rm-ANOVA in the PD group with the within-subject factor medication (ie, before, after, and at best) (Equation S4 in Multimedia Appendix 1). Participants with PD who had at least one marked task for each of the 3 PD medication conditions (ie, before, after, and at best) were included in treatment effect analysis (188 for gait, 189 for balance, 280 for voice, and 338 for tapping).

Results

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. Overall, 66 out of 423 gait, 59 out of 183 balance, 60 out of 124 voice, and 25 out of 43 tapping features differed significantly (all *P*s<.05) between PD and HC at baseline (Figure 1C) 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 2 and Tables S5-S8 in Multimedia Appendix 1).

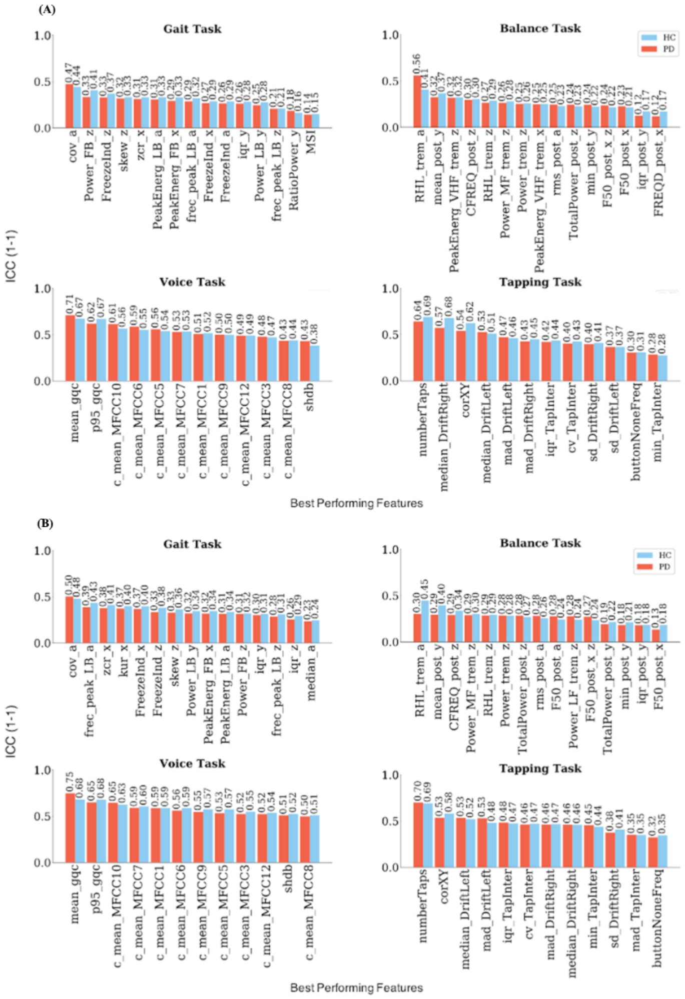
Figure 2. Effect size (Cohen d) for the most reliable features in the Parkinson disease and healthy control groups selected from different time points and repetitions. a: accelerometer average signal; iqr: interquartile range; min: minimum value; PeakEnerg: peak of energy; x: accelerometer mediolateral signal; y: accelerometer vertical signal; z: accelerometer anteroposterior signal. (A) Gait task. cov: coefficient of variation; FB: freezing band; frec_peak: frequency at the peak of energy; FreezeInd: freeze index; kur: kurtosis; LB: locomotor band; MSI: mean stride interval; RatioPower: sum of the power in the freezing and locomotor band; skew: skewness; zcr: zero-crossing rate. (B) Balance task. buttonNoneFreq: frequency of tapping outside the button; CFREQ: centroidal frequency; F50: frequency containing 50% of total power; FRQD: frequency of dispersion of the power spectrum; HF: high frequency (>4 Hz); LF: low frequency (0.15-3.5 Hz); MF: medium frequency (4-7 Hz); post: postural; Power: energy between 3.5-15 Hz; RHL: ratio between power in high frequency and low frequency; rms: root mean square; TotalPower: energy between 15-3.5 Hz; trem: tremor; VHF: very high frequency (>7 Hz). (C) Voice task. c_mean: mean of the MFCC; gqc: glottis quotient close; log: energy of the signal and the first and second derivatives of the MFCC; MFCC: Mel-frequency cepstral coefficients; p95: 95th percentile; shbd: shimmer. (D) Tapping task. corXY: correlation of X and Y positions; cv: coefficient; DriftLeft: left drift; DriftRight: right drift; mad: median absolute deviation; numberTaps: number of taps; sd: standard deviation; TapInter: tap interval.



Test-Retest Reliability

Next, we identified the top 10 features with highest median test-retest reliability (as measured using ICC) separately for PD and HC across different time points (one hour, one day, one week, or one month apart) and repetitions (all participants with 5 repetitions of the task) (Tables S5-S8 in Multimedia Appendix 1, Figure 1B). This procedure resulted in 12 to 15 features (including shared ones) being selected for each task (Figure 3, Figures S1 and S2 in Multimedia Appendix 1). 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 3). 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 0.11 to 0.09 for gait, from 0.21 to 0.13 for balance, from 0.39 to 0.24 for voice, and from 0.3 to 0.23 for tapping. The average ICC across the best performing features selected from different time points was also lower at one month compared to one hour apart, decreasing from 0.13 to 0.07 for gait, from 0.2 to 0.12 for balance, from 0.33 to 0.26 for voice, and from 0.32 to 0.19 for tapping.

Figure 3. Median ICC values for the most reliable features in the Parkinson disease and healthy control groups. a: accelerometer average signal; ICC: intraclass correlation coefficient; iqr: interquartile range; min: minimum value; PeakEnerg: peak of energy; x: accelerometer mediolateral signal; y: accelerometer vertical signal; z: accelerometer anteroposterior signal. (A) Median ICC values across different time points for the best performing features. (B) Median ICC values across different repetitions for the best performing features. Gait task—cov: coefficient of variation; FB: freezing band; frec_peak: frequency at the peak of energy; FreezeInd: freeze index; kur: kurtosis; LB: locomotor band; MSI: mean stride interval; RatioPower: sum of the power in the freezing and locomotor band; skew: skewness; zcr: zero-crossing rate. Balance task—buttonNoneFreq: frequency of tapping outside the button; CFREQ: centroidal frequency; F50: frequency containing 50% of total power; FRQD: frequency of dispersion of the power spectrum; HF: high frequency (>4 Hz); LF: low frequency (0.15-3.5 Hz); MF: medium frequency (4-7 Hz); post: postural; Power: energy between 3.5-15 Hz; RHL: ratio between power in high frequency and low frequency; rms: root mean square; TotalPower: energy between 15-3.5 Hz; trem: tremor; VHF: very high frequency (>7 Hz). Voice task—c_mean: mean of the MFCC; gqc: glottis quotient close; log: energy of the signal and the first and second derivatives of the MFCC; MFCC: Mel-frequency cepstral coefficients; p95: 95th percentile; shbd: shimmer. Tapping task—corXY: correlation of X and Y positions; cv: coefficient; DriftLeft: left drift; DriftRight: right drift; mad: median absolute deviation; numberTaps: number of taps; sd: standard deviation; TapInter: tap interval.



Repetition Effects

Next, we evaluated the longitudinal stability of these most reliable features. Using rm-ANOVA, we tested for the main effects of diagnosis, repetition (first, second, third, fourth, and fifth), and their interaction (Figures 4 and 5, Tables S9 and S10-S13 in Multimedia Appendix 1). 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, and 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. Further, we tested for the main effects of the elapsed time between repetitions and its interaction with diagnosis (Tables S18-S21 in Multimedia Appendix 1). A significant main effect of elapsed time was observed for 1 out of 15 gait features, 2 out of 15 balance features, 5 out of 12 voice features, and 5 out of 12 tapping features. A significant diagnosis-by-time interaction effect was observed only in 1 out of 15 balance features and 3 out of 12 tapping features.

Figure 4. Mean value of the best performing baseline features across different time points, calculated for PD and HC separately. a: accelerometer average signal; HC: healthy controls; iqr: interquartile range; min: minimum value; PD: Parkinson disease; PeakEnerg: peak of energy; x: accelerometer mediolateral signal; y: accelerometer vertical signal; z: accelerometer anteroposterior signal. (A) Gait task. cov: coefficient of variation; FB: freezing band; frec_peak: frequency at the peak of energy; FreezeInd: freeze index; LB: locomotor band; MSI: mean stride interval; RatioPower: sum of the power in the freezing and locomotor band; skew: skewness; zcr: zero-crossing rate. (B) Balance task. buttonNoneFreq: frequency of tapping outside the button; CFREQ: centroidal frequency; F50: frequency containing 50% of total power; FRQD: frequency of dispersion of the power spectrum; HF: high frequency (>4 Hz); LF: low frequency (0.15-3.5 Hz); MF: medium frequency (4-7 Hz); post: postural; Power: energy between 3.5-15 Hz; RHL: ratio between power in high frequency and low frequency; rms: root mean square; TotalPower: energy between 15-3.5 Hz; trem: tremor; VHF: very high frequency (>7 Hz). (C) Voice task. c_mean: mean of the MFCC; gqc: glottis quotient close; log: energy of the signal and the first and second derivatives of the MFCC; MFCC: Mel-frequency cepstral coefficients; p95: 95th percentile; shbd: shimmer. (D) Tapping task. corXY: correlation of X and Y positions; cv: coefficient; DriftLeft: left drift; DriftRight: right drift; mad: median absolute deviation; numberTaps: number of taps; sd: standard deviation; TapInter: tap interval.

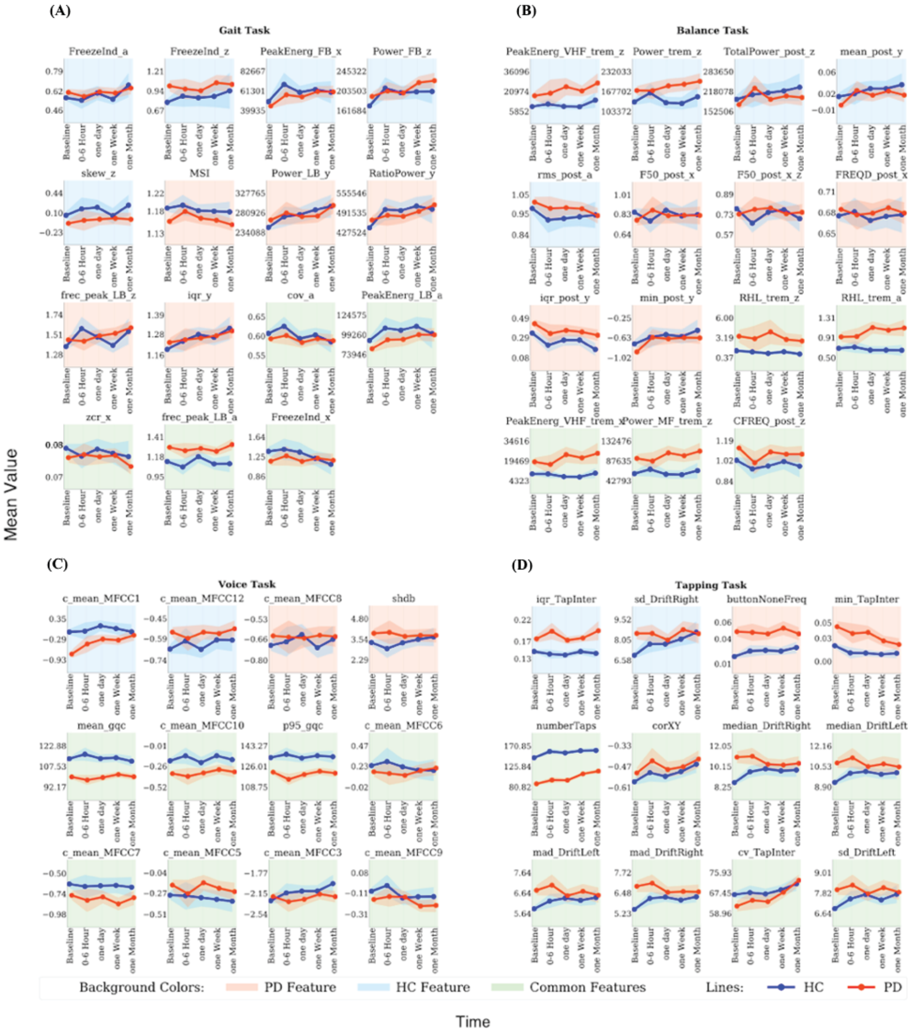
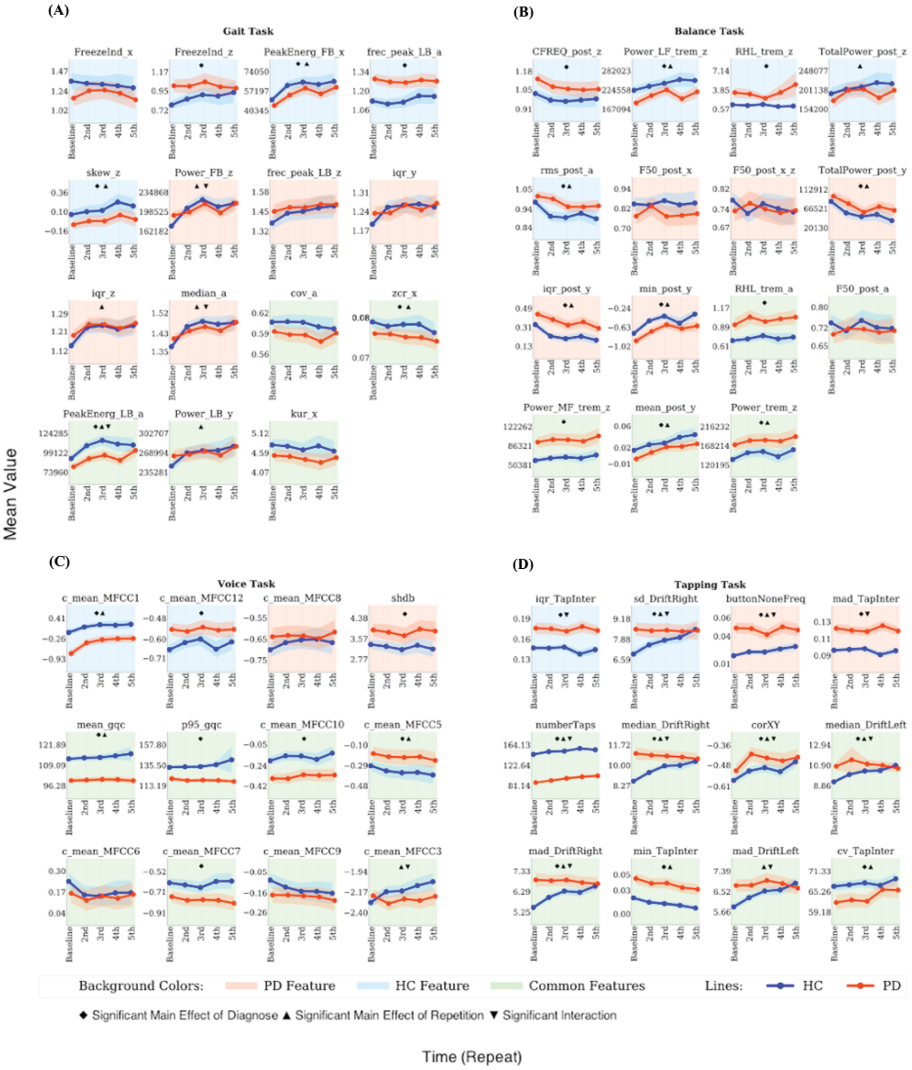


Figure 5. Mean value of the best performing baseline features across repetitions, calculated for PD and HC separately. a: accelerometer average signal; HC: healthy controls; iqr: interquartile range; min: minimum value; PD: Parkinson disease; PeakEnerg: peak of energy; x: accelerometer mediolateral signal; y: accelerometer vertical signal; z: accelerometer anteroposterior signal. (A) Gait task. cov: coefficient of variation; FB: freezing band; frec_peak: frequency at the peak of energy; FreezeInd: freeze index; kur: kurtosis; LB: locomotor band; MSI: mean stride interval; RatioPower: sum of the power in the freezing and locomotor band; skew: skewness; zcr: zero-crossing rate. (B) Balance task. buttonNoneFreq: frequency of tapping outside the button; CFREQ: centroidal frequency; F50: frequency containing 50% of total power; FRQD: frequency of dispersion of the power spectrum; HF: high frequency (>4 Hz); LF: low frequency (0.15-3.5 Hz); MF: medium frequency (4-7 Hz); post: postural; Power: energy between 3.5-15 Hz; RHL: ratio between power in high frequency and low frequency; rms: root mean square; TotalPower: energy between 15-3.5 Hz; trem: tremor; VHF: very high frequency (>7 Hz). (C) Voice task. c_mean: mean of the MFCC; gqc: glottis quotient close; log: energy of the signal and the first and second derivatives of the MFCC; MFCC: Mel-frequency cepstral coefficients; p95: 95th percentile; shbd: shimmer. (D) Tapping task. corXY: correlation of X and Y positions; cv: coefficient; DriftLeft: left drift; DriftRight: right drift; mad: median absolute deviation; numberTaps: number of taps; sd: standard deviation; TapInter: tap interval.



In an additional sensitivity analysis, we further tested if the between-group differences and group-by-repetition interaction remain significant when controlling for age and sex. The results (Tables S14-S17 in Multimedia Appendix 1) show that a significant effect of diagnosis was still identified for 2 out of 6 gait features, 8 out of 11 balance features, 1 out of 8 voice features, and 10 out of 11 tapping features. A significant effect of repetition was still found for 6 out of 8 gait features, 7 out of 8 balance features, 3 out of 4 voice features, and 10 out of 10 tapping features. Also, a significant main effect of diagnosis-by-repetition was still observed for 1 out of 3 gait features, 1 out of 1 balance feature, and 8 out of 10 tapping features.

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, Tables S9 and S10-S13, medication column, in Multimedia Appendix 1).

Discussion

Principal Findings

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 a few features had medium to large effect sizes for differentiation between PD and HC. For all tasks, a substantial percentage of features displayed significant longitudinal alterations in their mean values over time.

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. Balance and gait tasks displayed consistently poor test-retest reliabilities as well as low effect sizes for differentiation between PD and HC, calling into question 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.

Unlike some previous studies that showed good performance and moderate to excellent correlation of gait and balance features with clinical score [4,39], the overall poor performance of these tasks in the m-Power study may be explained by the nature of these tasks, which requires strict supervision and monitoring. Both may not be sufficiently achieved in the self-administered setting of the m-Power study. Overall, acceleration-related features in the gait task and tremor-related features and those selected from frequency domain in the balance task displayed the best performance for the respective task [23,40]. The features related to Mel-frequency cepstral coefficients for the voice task displayed the highest effect sizes for this task, which is in line with previous studies showing its ability in identifying pathological speech [41,42]. In line with previous studies, features related to intertapping interval and precision of the tapping task (eg, number of taps, taps drift) displayed the best performance among all [43,44].

Most features showed a decrease in test-retest reliability with longer periods of time. This may reflect a consequence of the repetition effects and the group-by-repetition interaction observed in the analyses of variance for a substantial proportion of the features. Features selected from the tapping task were less sensitive to the effect of age and sex compared to other tasks. Overall, the effects of age and sex were not significant for most of the features. The analysis of elapsed time between repetitions also revealed that the time difference between repetitions did not have a significant effect on most of the features. ICC values obtained from the PD and HC groups were largely similar, suggesting that other non-PD related sources of variation may have played a larger role in the observed low ICC values. Determining these reasons requires more controlled experiments than provided by the m-Power study.

Despite a 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 [2,10,12]. While a clear differentiation of motivation versus 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 toward PD is likely to reflect worsening due to reduced motivation, disease progression, or other similar factors. In contrast, movements toward 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 [45-47]. 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. While differential learning has been previously reported [45,48-52], 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 that 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 a 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 [20], might be the self-reported nature of the medication status in the m-Power data set, which likely introduced some noise variation (ie, 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 data set is also the major limitation of our study. In such an uncontrolled setting, accuracy in reporting the diagnosis and demographics, defining the medication status, and ensuring correct understanding of 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. Another limitation of our study is the moderate adherence of participants in the m-Power study, which limited the number of participants who could be included in our analyses. Differences in age as well as lack of standardization of the time of day when the assessments were conducted are further sources of variation that may affect the generalizability of our findings [53]. Future studies may make inferences about the impact of different confounders such as comorbidities and disease severity on the longitudinal stability of DB. Also, further research is needed to establish the longitudinal stability of DB in the context of their relationship to clinical rating scales such as UPDRS.

Nonetheless, our findings clearly demonstrate the need for further optimization of DB tasks as well as for introducing careful monitoring and quality control procedures to enable integration of DB measures into clinically relevant applications.

Data Availability

The m-Power data set used for this paper is available upon registration from Synapse [54].

Acknowledgments

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 785907 (Human Brain Project SGA2).

The data used in this study were contributed by users of the Parkinson m-Power mobile app as part of the m-Power study developed by Sage Bionetworks and described in Synapse [54].

Authors' Contributions

MSF performed analyses and wrote the 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.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplement.
[DOCX File , 1606 KB-Multimedia Appendix 1]

References

1. Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008 Apr;79(4):368-376. [doi: 10.1136/jnnp.2007.131045] [Medline: 18344392]
2. Prince J, Arora S, de Vos M. Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. Physiol Meas 2018 Apr 26;39(4):044005. [doi: 10.1088/1361-6579/aab512] [Medline: 29516871]
3. Insel TR. Digital Phenotyping: Technology for a New Science of Behavior. JAMA 2017 Oct 03;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
4. Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord 2018 Aug;33(8):1287-1297 [FREE Full text] [doi: 10.1002/mds.27376] [Medline: 29701258]
5. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med 2019;2(1):14 [FREE Full text] [doi: 10.1038/s41746-019-0090-4] [Medline: 30868107]
6. Mahadevan N, Demanuele C, Zhang H, Volfson D, Ho B, Erb MK, et al. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. NPJ Digit Med 2020;3:5 [FREE Full text] [doi: 10.1038/s41746-019-0217-7] [Medline: 31970290]
7. Shah VV, McNames J, Mancini M, Carlson-Kuhta P, Nutt JG, El-Gohary M, et al. Digital Biomarkers of Mobility in Parkinson's Disease During Daily Living. J Parkinsons Dis 2020;10(3):1099-1111 [FREE Full text] [doi: 10.3233/JPD-201914] [Medline: 32417795]

8. Schlachetzki JCM, Barth J, Marxreiter F, Gossler J, Kohl Z, Reinfelder S, et al. Wearable sensors objectively measure gait parameters in Parkinson's disease. PLoS One 2017;12(10):e0183989 [FREE Full text] [doi: 10.1371/journal.pone.0183989] [Medline: 29020012]
9. Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform 2020 Apr;104:103362 [FREE Full text] [doi: 10.1016/j.jbi.2019.103362] [Medline: 31866434]
10. Espay AJ, Bonato P, Nahab FB, Maetzler W, Dean JM, Klucken J, Movement Disorders Society Task Force on Technology. Technology in Parkinson's disease: Challenges and opportunities. Mov Disord 2016 Sep;31(9):1272-1282 [FREE Full text] [doi: 10.1002/mds.26642] [Medline: 27125836]
11. Moore ST, Yungher DA, Morris TR, Dilda V, MacDougall HG, Shine JM, et al. Autonomous identification of freezing of gait in Parkinson's disease from lower-body segmental accelerometry. J Neuroeng Rehabil 2013 Feb 13;10:19 [FREE Full text] [doi: 10.1186/1743-0003-10-19] [Medline: 23405951]
12. Fisher JM, Hammerla NY, Rochester L, Andras P, Walker RW. Body-Worn Sensors in Parkinson's Disease: Evaluating Their Acceptability to Patients. Telemed J E Health 2016 Jan;22(1):63-69 [FREE Full text] [doi: 10.1089/tmj.2015.0026] [Medline: 26186307]
13. Rahlf AL, 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 2019;1:27 [FREE Full text] [doi: 10.3389/fspor.2019.00027] [Medline: 33344951]
14. Orlowski K, Eckardt F, Herold F, Aye N, Edelmann-Nusser J, Witte K. Examination of the reliability of an inertial sensor-based gait analysis system. Biomed Tech (Berl) 2017 Nov 27;62(6):615-622. [doi: 10.1515/bmt-2016-0067] [Medline: 28099115]
15. Hasegawa N, Shah VV, Carlson-Kuhta P, Nutt JG, Horak FB, Mancini M. How to Select Balance Measures Sensitive to Parkinson's Disease from Body-Worn Inertial Sensors-Separating the Trees from the Forest. Sensors (Basel) 2019 Jul 28;19(15):3320 [FREE Full text] [doi: 10.3390/s19153320] [Medline: 31357742]
16. 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;2013:389195 [FREE Full text] [doi: 10.1155/2013/389195] [Medline: 24386590]
17. 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 2017;9:11-17. [doi: 10.1016/j.imu.2017.05.005]
18. Wissel B, Mitsi G, Dwivedi A, Papapetropoulos S, Larkin S, López Castellanos JR, et al. Tablet-Based Application for Objective Measurement of Motor Fluctuations in Parkinson Disease. Digit Biomark 2017;1(2):126-135 [FREE Full text] [doi: 10.1159/000485468] [Medline: 32095754]
19. Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data 2016 Mar 03;3:160011 [FREE Full text] [doi: 10.1038/sdata.2016.11] [Medline: 26938265]
20. Zhan A, Little M, Harris D, Abiola S, Dorsey E, Saria S. High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection. arXiv. Preprint posted online on January 5, 2016 [FREE Full text]
21. Schwab P, Karlen W. PhoneMD: Learning to Diagnose Parkinson's Disease from Smartphone Data. 2019 Jul 17 Presented at: Thirty-Third AAAI Conference on Artificial Intelligence; January 27-February 1, 2019; Honolulu, HI p. 1118-1125. [doi: 10.1609/aaai.v33i01.33011118]
22. Arora S, Venkataraman V, Zhan A, Donohue S, Biglan KM, Dorsey ER, et al. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. Parkinsonism Relat Disord 2015 Jun;21(6):650-653. [doi: 10.1016/j.parkreldis.2015.02.026] [Medline: 25819808]
23. Goñi M, Eickhoff S, Far M, Patil K, Dukart J. Limited diagnostic accuracy of smartphone-based digital biomarkers for Parkinson's disease in a remotely-administered setting. medRxiv. Preprint posted online on January 15, 2021 [FREE Full text] [doi: 10.1101/2021.01.13.21249660]
24. Mirelman A, Heman T, Yasinovsky K, Thaler A, Gurevich T, Marder K, LRRK2 Ashkenazi Jewish Consortium. Fall risk and gait in Parkinson's disease: the role of the LRRK2 G2019S mutation. Mov Disord 2013 Oct;28(12):1683-1690. [doi: 10.1002/mds.25587] [Medline: 24123150]
25. Blin O, Ferrandez AM, Serratrice G. Quantitative analysis of gait in Parkinson patients: increased variability of stride length. J Neurol Sci 1990 Aug;98(1):91-97. [doi: 10.1016/0022-510x(90)90184-o] [Medline: 2230833]
26. Hausdorff JM. Gait dynamics in Parkinson's disease: common and distinct behavior among stride length, gait variability, and fractal-like scaling. Chaos 2009 Jun;19(2):026113 [FREE Full text] [doi: 10.1063/1.3147408] [Medline: 19566273]
27. Miller Koop M, Ozinga SJ, Rosenfeldt AB, Alberts JL. Quantifying turning behavior and gait in Parkinson's disease using mobile technology. IBRO Rep 2018 Dec;5:10-16 [FREE Full text] [doi: 10.1016/j.ibror.2018.06.002] [Medline: 30135951]
28. 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 2012;15(9):899-910. [doi: 10.1080/10255842.2011.565753] [Medline: 21547782]

29. Prieto TE, Myklebust JB, Hoffmann RG, Lovett EG, Myklebust BM. Measures of postural steadiness: differences between healthy young and elderly adults. IEEE Trans Biomed Eng 1996 Sep;43(9):956-966. [doi: 10.1109/10.532130] [Medline: 9214811]
30. Palakurthi B, Burugupally SP. Postural Instability in Parkinson's Disease: A Review. Brain Sci 2019 Sep 18;9(9):239 [FREE Full text] [doi: 10.3390/brainsci9090239] [Medline: 31540441]
31. Pittman B, Ghomi RH, Si D. Parkinson's Disease Classification of mPower Walking Activity Participants. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:4253-4256. [doi: 10.1109/EMBC.2018.8513409] [Medline: 30441293]
32. Seifert K, Camacho O. Implementing Positioning Algorithms Using Accelerometers. 2007 Feb. URL: https://www.nxp.com/docs/en/application-note/AN3397.pdf [accessed 2021-02-01]
33. 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 2011 May;15(3):481-490. [doi: 10.1109/TITB.2011.2107916] [Medline: 21349795]
34. Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online 2007 Jun 26;6:23 [FREE Full text] [doi: 10.1186/1475-925X-6-23] [Medline: 17594480]
35. 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 2020 Jun 01;70(11):393-405 [FREE Full text] [doi: 10.33588/rn.7011.2019414] [Medline: 32436206]
36. Rao G, Fisch L, Srinivasan S, D'Amico F, Okada T, Eaton C, et al. Does this patient have Parkinson disease? JAMA 2003 Jan 15;289(3):347-353. [doi: 10.1001/jama.289.3.347] [Medline: 12525236]
37. Jobbágy A, Harcos P, Karoly R, Fazekas G. Analysis of finger-tapping movement. J Neurosci Methods 2005 Jan 30;141(1):29-39. [doi: 10.1016/j.jneumeth.2004.05.009] [Medline: 15585286]
38. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment 1994 Dec;6(4):284-290. [doi: 10.1037/1040-3590.6.4.284]
39. Horak FB, Mancini M. Objective biomarkers of balance and gait for Parkinson's disease using body-worn sensors. Mov Disord 2013 Sep 15;28(11):1544-1551 [FREE Full text] [doi: 10.1002/mds.25684] [Medline: 24132842]
40. Sejdić E, Lowry KA, Bellanca J, Redfern MS, Brach JS. A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains. IEEE Trans Neural Syst Rehabil Eng 2014 May;22(3):603-612 [FREE Full text] [doi: 10.1109/TNSRE.2013.2265887] [Medline: 23751971]
41. Khan T. Running-speech MFCC are better markers of Parkinsonian speech deficits than vowel phonation and diadochokinetic. DiVA. Preprint posted online on July 15, 2015 [FREE Full text]
42. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng 2012 May;59(5):1264-1271. [doi: 10.1109/TBME.2012.2183367] [Medline: 22249592]
43. Memedi M, Khan T, Grenholm P, Nyholm D, Westin J. Automatic and objective assessment of alternating tapping performance in Parkinson's disease. Sensors (Basel) 2013 Dec 09;13(12):16965-16984 [FREE Full text] [doi: 10.3390/s131216965] [Medline: 24351667]
44. Lee CY, Kang SJ, Hong S, Ma H, Lee U, Kim YJ. A Validation Study of a Smartphone-Based Finger Tapping Application for Quantitative Assessment of Bradykinesia in Parkinson's Disease. PLoS One 2016;11(7):e0158852 [FREE Full text] [doi: 10.1371/journal.pone.0158852] [Medline: 27467066]
45. Olson M, Lockhart TE, 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 2019;10:62 [FREE Full text] [doi: 10.3389/fneur.2019.00062] [Medline: 30792688]
46. Steib S, Wanner P, Adler W, Winkler J, Klucken J, Pfeifer K. A Single Bout of Aerobic Exercise Improves Motor Skill Consolidation in Parkinson's Disease. Front Aging Neurosci 2018;10:328 [FREE Full text] [doi: 10.3389/fnagi.2018.00328] [Medline: 30405397]
47. Bryant M, Workman CD, Jamal F, Meng H, Jackson GR. Feasibility study: Effect of hand resistance exercise on handwriting in Parkinson's disease and essential tremor. J Hand Ther 2018;31(1):29-34. [doi: 10.1016/j.jht.2017.01.002] [Medline: 28389133]
48. Krebs HI, Hogan N, Hening W, Adamovich SV, Poizner H. Procedural motor learning in Parkinson's disease. Exp Brain Res 2001 Dec;141(4):425-437. [doi: 10.1007/s002210100871] [Medline: 11810137]
49. Sehm B, Taubert M, Conde V, Weise D, Classen J, Dukart J, et al. Structural brain plasticity in Parkinson's disease induced by balance training. Neurobiol Aging 2014 Jan;35(1):232-239. [doi: 10.1016/j.neurobiolaging.2013.06.021] [Medline: 23916062]
50. Foreman KB, Sondrup S, Dromey C, Jarvis E, Nissen S, Dibble LE. The effects of practice on the concurrent performance of a speech and postural task in persons with Parkinson disease and healthy controls. Parkinsons Dis 2013;2013:987621 [FREE Full text] [doi: 10.1155/2013/987621] [Medline: 23841022]

51. Agostino R, Currà A, Soldati G, Dinapoli L, Chiacchiari L, Modugno N, et al. Prolonged practice is of scarce benefit in improving motor performance in Parkinson's disease. Mov Disord 2004 Nov;19(11):1285-1293. [doi: 10.1002/mds.20247] [Medline: 15390058]
52. Behrman AL, Cauraugh JH, Light KE. Practice as an intervention to improve speeded motor performance and motor learning in Parkinson's disease. J Neurol Sci 2000 Mar 15;174(2):127-136. [doi: 10.1016/s0022-510x(00)00267-7] [Medline: 10727698]
53. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3:21 [FREE Full text] [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
54. mPower Public Researcher Portal. URL: https://www.synapse.org/#!Synapse:syn4993293 [accessed 2021-02-01]

Abbreviations

DB: digital biomarkers
HC: healthy controls
ICC: intraclass correlation coefficients
PD: Parkinson disease
rm-ANOVA: repeated-measures analyses of variance
UPDRS: Unified Parkinson's Disease Rating Scale

Edited by R Kukafka; submitted 29.12.20; peer-reviewed by M Lindemann, O Chen, O Beiki, Z Feng, Y Cai; comments to author 03.03.21; revised version received 21.06.21; accepted 05.07.21; published 13.09.21

Please cite as:
Sahandi Far M, Eickhoff SB, Goni M, Dukart J
Exploring Test-Retest Reliability and Longitudinal Stability of Digital Biomarkers for Parkinson Disease in the m-Power Data Set: Cohort Study
J Med Internet Res 2021;23(9):e26608
URL: https://www.jmir.org/2021/9/e26608
doi: 10.2196/26608
PMID:

Supplementary Materials for [Sahandi-Far et al].

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

Equation S1. $y \sim diagosis * repetition + (1|subject)$

Equation S2. $y \sim diagosis * repetition + age + sex + (1|subject)$

Equation S3. $y \sim diagosis * eleapse\_time + age + sex + (1|subject)$

Equation S4. $y \sim medication + (1|subject)$

Where y is the response.

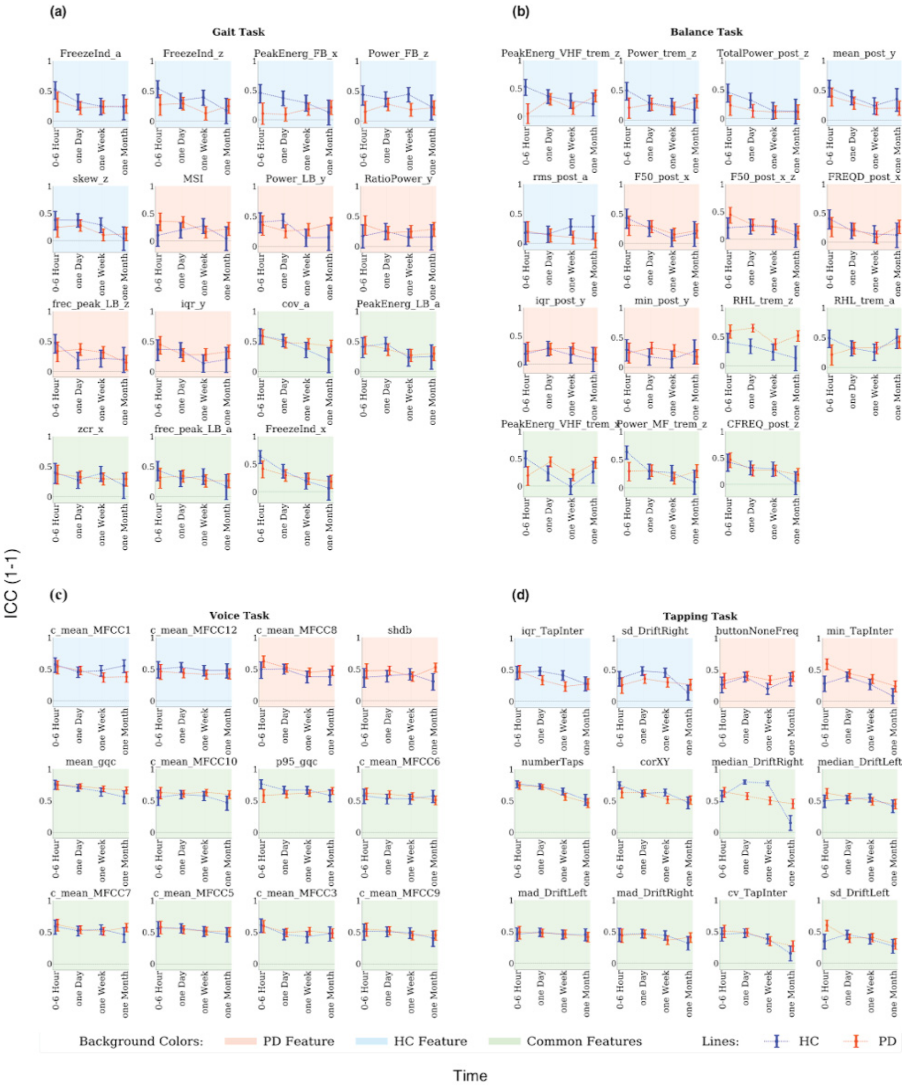


Figure S1. Test-retest reliability for the most reliable features in PD and HC selected from different time points. Intraclass Correlation Coefficient ICC (1-1) values with a 95% confidence interval across different time points vs baseline. **(a)** Gait Task: FreezeInd- Freeze Index, PeakEnerg- Peak 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, **(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 Absolute Deviation, min- Minimum, cv- Coefficient, Sd- Standard Deviation.

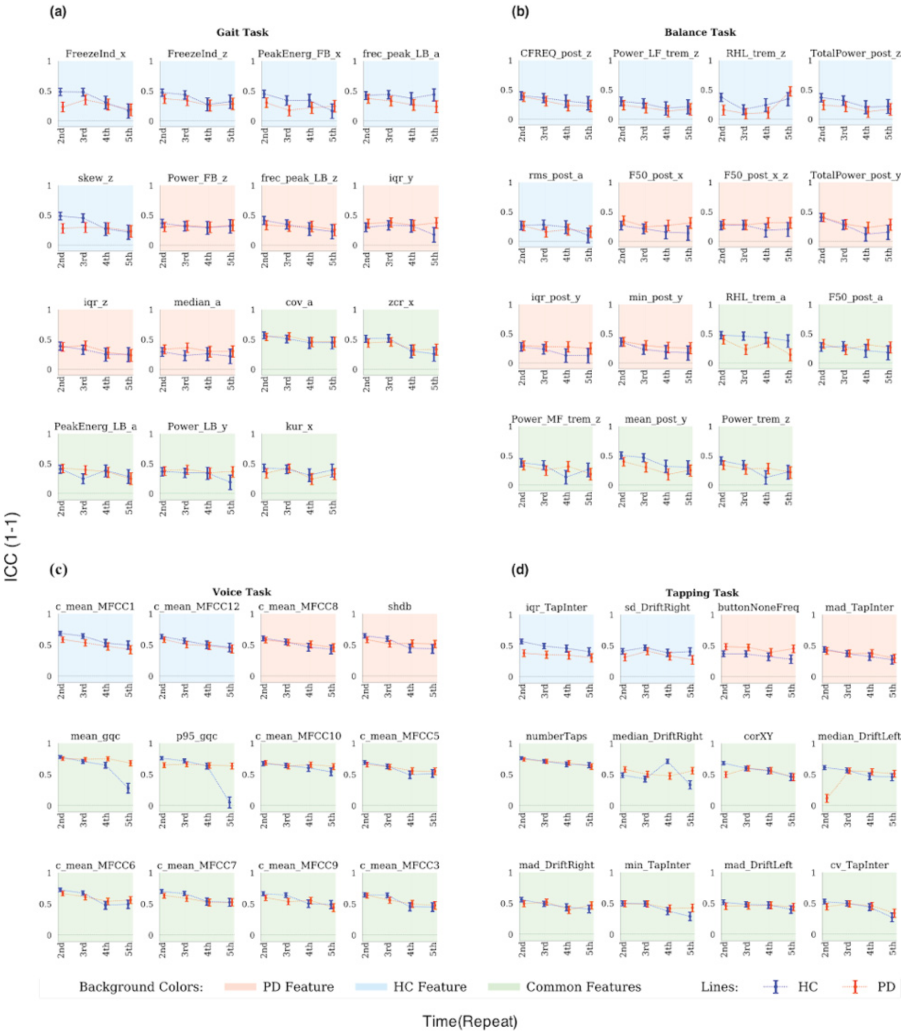


Figure S2. Test-retest reliability for the most reliable features in PD and HC selected from different repetitions. Intraclass Correlation Coefficient ICC (1-1) values with a 95% confidence interval across different repetition vs baseline. **(a)** Walking Task: FreezeInd- Freeze Index, PeakEnerg- Peak 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 Absolute Deviation, min- Minimum, cv- Coefficient, Sd- Standard Deviation.

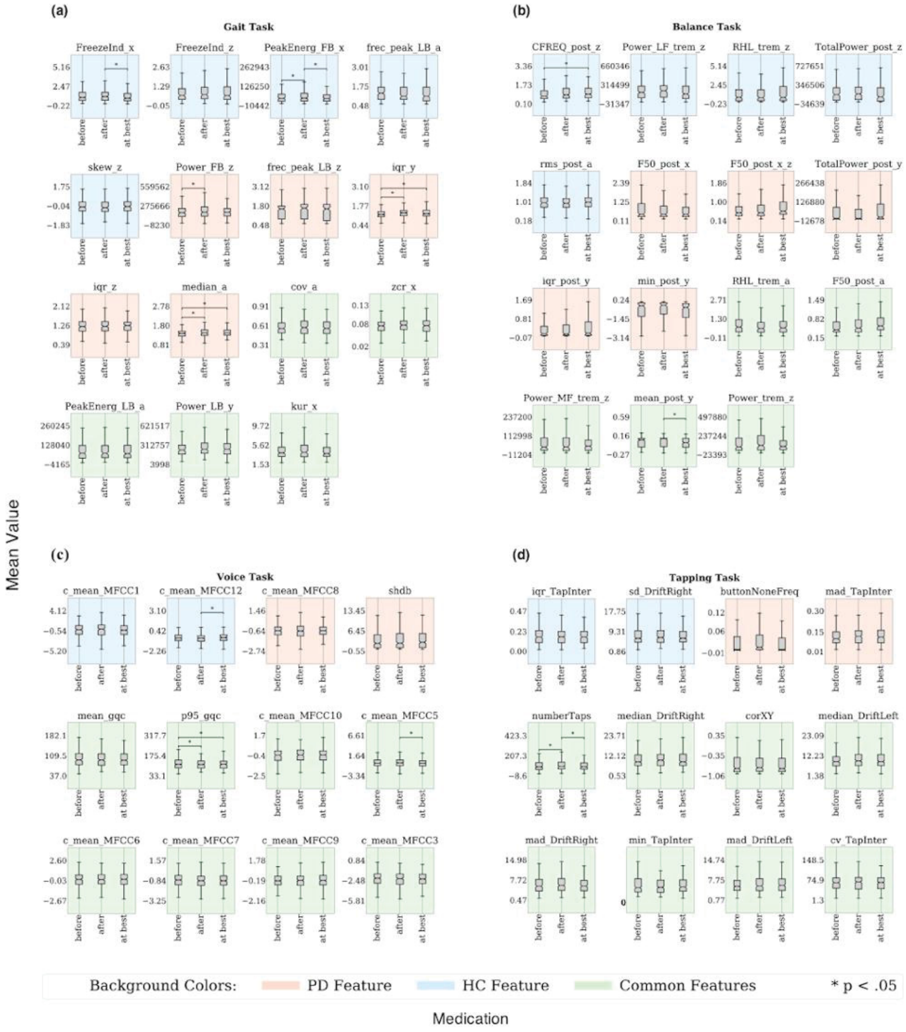


Figure S3. Mean Value of the best performing features at different medication conditions. **(a)** Gait Task: FreezeInd- Freeze Index, PeakEnerg- Peak of Energy, frec_peak- Frequency at the Peak of Energy, skew- Skewness, iqr- Interquartile Range, cov- Coefficient of Variation, zcr- Zero-Crossing Rate, 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 Absolute Deviation, min- Minimum, cv- Coefficient, Sd- Standard Deviation.

Table S1. Gait Features.

Feature acronym	Units	Feature description	Signal (acronym)
numSteps		Number of steps during the 10 seconds gait signal.	
MSI	s	Mean Stride Interval, calculated as the duration of a stride averaged over all strides[1,2].	
StrideVar	%	Stride Variability, calculated as the standard deviation divided by the mean stride of the stride interval. Measures consistency and stability[1,2].	
mean	ms^-2	Mean value of the observations[3,4].	Mediolateral, vertical, anteroposterior, and average acceleration (x, y, z, a) Mediolateral, vertical, anteroposterior, and average postural acceleration (post_x, post _y, post _z, post_a) velocity (vel) acceleration along path (acc)
min	ms^-2	Minimum value of the observations.	
max	ms^-2	Maximum value of the observations.	
median	ms^-2	Median. Middle value among a dataset[3,4].	
sd	ms^-2	Standard deviation, calculated as the sum of squares differences between the individual values and the mean. Measures variability[3,4].	
var	(ms^2)^2	Variance, calculated as the square of the standard deviation. Measures variability.	
range	ms^-2	Range of the observations.	
iqr	ms^-2	Interquartile range, calculated as the difference between 75th and 25th percentiles. Measures dispersion[3,4].	
rms	ms^-2	Root mean square of the observations.	
cov		Coefficient of variation, calculated as the standard deviation of the signal divided by the mean.	
skew		Skewness. Describes the asymmetry of a signal. A negative value indicates that the distribution is concentrated on the right, while a positive one is concentrated in the left[2–4].	
kur		Kurtosis. Measures if data is heavy or light-tailed to a normal distribution[2,3].	
zcr		Zero-crossing rate. Rates sign-changes along a signal[4].	

ApEn		Entropy. Measures uncertainty, ranging from 0-1 where 0 indicates randomness and 1 maximum regularity[2,4].	
PeakEnerg_LB	psd	Peak of energy in the locomotor band (0.5-3 Hz)[1,5].	
frec_peak_LB	Hz	Frequency at the peak of energy in the locomotor band (0.5-3 Hz)[5].	
Power_LB	psd	Power of the locomotor band (0.5-3 Hz)[5].	
PeakEnerg_FB	psd	Peak of energy in the freezing band (3-8Hz)[6].	
frec_peak_FB	Hz	Frequency at the peak of energy in the freezing band (3-8 Hz).	
Power_FB	psd	Power in the freezing band (3-8 Hz).	
FreezeInd		Freeze Index. Calculated as the ratio between the power in the freezing band (3-8 Hz) and the power in the locomotor band (0.5-3 Hz)[6].	
RatioPower	psd	Sum of the power in the freezing (3-8 Hz) and locomotor band (3-8 Hz)[7]	
ar		Coefficient of a 1st order autoregressive model. An autoregressive model forecasts when there is some correlation between current values and their preceding ones.	
COEFCEPS_(1-20)	mel	20 Mel Frequency Cepstral Coefficients. Represent the short-term power spectrum[6]	

Table S2. Balance Features.

Acronym	units	Description	Signal (acronym)
mean	ms^-2	Mean value of the observations.	Mediolateral, vertical, anteroposterior, and average tremor acceleration (trem_x, trem_y, trem_z, trem_a)
min	ms^-2	Minimum value of the observations.	
max	ms^-2	Maximum value of the observations.	
median	ms^-2	Median value of the observations.	
sd	ms^-2	Standard deviation, calculated as the sum of squares differences between the individual values and the mean. Measures variability.	
var	mg^2	Variance, calculated as the square of the standard deviation. Measures variability.	Mediolateral, vertical, anteroposterior, and average postural acceleration (post_x, post _y, post _z, post_a)
range	ms^-2	Range of the observations.	
iqr	ms^-2	Interquartile range, calculated as the difference between 75th and 25th percentiles. Measures dispersion.	
rms	ms^-2	Root mean square of the observations.	
cov	ms^-2	Coefficient of variation, calculated as the standard deviation of the signal divided by the mean.	
skew		Skewness. Describes the asymmetry of a signal. A negative value indicates that the distribution is concentrated on the right,	

		while a positive one is concentrated in the left.	
kur		Kurtosis. Measures if data is heavy or light-tailed to a normal distribution.	
zcr		Zero-crossing rate. Rates sign-changes along a signal.	
ApEn		Entropy. Measures uncertainty, ranging from 0-1 where 0 indicates randomness and 1 maximum regularity.	
Power_MF	psd	Power of the medium frequency band 4-7Hz [8].	Mediolateral, vertical, anteroposterior, and average tremor acceleration (trem_x, trem_y, trem_z, trem_a)
PeakEnerg_VHF	psd	Peak of energy in the very high frequency band (>7Hz)	
frec_peak_HF	Hz	Frequency at the peak of energy in high frequency band (>4Hz)[8].	
Power	psd	Power between 3.5-15Hz	
Power_LF	psd	Power in the low frequency band (0.15-3.5Hz)	
RHL		Ratio between the power between 3.5-15Hz and power between 0.15-3.5Hz[8].	
CFREQ	Hz	Centroidal frequency for postural measures. Also known as zero-crossing frequency[8–10].	Mediolateral, vertical, anteroposterior, and average postural acceleration (post_x, post _y, post _z, post_a) Mediolateral-anteroposterior average postural acceleration (post_x_z)
FREQD	Hz	Frequency of dispersion of the power spectrum for postural measures[8–10].	
jerk	m^2/s^2	Average jerk. Measures vibration as the rate of change in acceleration. Calculated as the derivative of acceleration with respect to time[8,9].	Mediolateral, vertical, anteroposterior, and average postural acceleration (post_x, post _y, post _z, post_a) Mediolateral-anteroposterior average postural acceleration (post_x_z)
TotalPower	pwd	Energy between 0.15-3.5Hz for postural measures[9].	
F50	Hz	Frequency containing 50% of the total power for postural measures[8,9].	
F95	Hz	Frequency containing 95% of the total power for postural measures[8,9].	
MDIST	mm	Represents the average distance from the center to each AP and ML points[8,10].	Mediolateral, anteroposterior and average of mediolateral
RDIST	mm	Root Mean Square distance from the mean center[9,10].	

TOTEX	mm	Total excursions is the total length of the path. Calculated as the sum of distances between consecutive points[10].	and anteroposterior distance (dist_x, dist_z,dist_x_z)
MVELO	mm/s	Mean velocity is the average velocity of the center path, calculated as the TOTEX divided by the time[8,10].	
MFREQ	mm/s	The mean frequency is the rotational frequency with a radius equal to the mean distance[9,10].	
AREA_CC	mm^2	The 95% confidence circle area is the area of a circle enclosing all points in the AP-ML plane with 95% confidence[9,10].	average of mediolateral and anteroposterior distance (dist_x_z)
AREA_CE	mm^2	The 95% confidence ellipse area is the area of an ellipse enclosing all points in the AP-ML plane with 95% confidence[8–10].	
AREA_SW	mm^2/s	Sway area calculated as the area enclosing the acceleration path[8–10].	
FD		The fractal dimension indicates the degree to which a curve fills the enclosed metric space[9,10].	
FD_CC		Fractal dimension based on the 95% confidence circle area[9,10].	
FD_CE		Fractal dimension based on the 95% confidence ellipse area[9,10].	

Table S3. Voice Features

acronym	Units	description	Signal (acronym)
amp	psd^0.5	Average amplitude[11].	
shim	%	Absolute shimmer[4,11,12].	
shdb	db	Shimmer in logarithmic domain[11].	
apq3	%	3 point amplitude perturbation quotient in percentage[11].	
apq5	%	5 point amplitude perturbation quotient in percentage[11].	
fm	Hz	Frequency modulation[11].	
hnr_mean	db	Mean of the harmonic to noise ratio, which indicates the amount of noise[4,11,12].	
hnr_std	db	Standard deviation of the harmonic to noise ratio[11].	
rpde		Recurrence period density entropy. Characterizes the deviation from signal periodicity[4,11].	
DFA		Detrended Fluctuation Analysis, which describes turbulent noise[4,11].	
mean		Mean value[4,11].	fundamental frequency (f0), amplitude (amp), Teager Kaiser Energy Operator of the
sd		Standard deviation[4,11].	

			fundamental frequency (tkeo), open quotient (oq), glottis quotient open (gqo), glottis quotient closed (gqc)
jitt		Absolute jitter[4,11].	fundamental frequency (f0), period (T)
jitta		Relative or local jitter[11].	
rap		Relative average perturbation[11].	
ppq5	St	Perturbation quotient using 5 point (cycles)[11].	
range		Range[11].	Teager Kaiser Energy Operator of the fundamental frequency (tkeo),
p25	Hz^2	25th percentile of the Teager-Kaiser Energy Operator[11].	
p75	Hz^2	75th percentile of the Teager-Kaiser Energy Operator[11].	
ApEn		Pitch Period Entropy. Quantifies the impaired control of stable pitch during a sustained phonation[4,11].	
p5	Hz^2 (teko)	5th percentile[11].	Teager Kaiser Energy Operator of the fundamental frequency (tkeo), open quotient (oq), glottis quotient open (gqo), glottis quotient closed (gqc)
p95	Hz^2	95th percentile[11].	
c_mean	db, mel, 1st and 2nd derivative of mel	Mean of the MFCCs coefficients, log-energy of the signal and the first and second derivatives of the MFCCs[4,11,12].	log energy (log), 0th order cepstral coefficient (0th), 1-12th Mel Frequency Cepstral Coefficients (MFCC_(1-12), 1-14th deltas (d_(1-14)), 1-14th delta-delta (dd_(1-14))
c_std	db, mel, 1st and 2nd derivative of mel	Standard deviation of the MFCCs coefficients, log-energy of the signal and the first and second derivatives of the MFCCs[11,12].	

Table S4. Tapping Features.

acronym	units	description	Signal (acronym)
numberTaps		Number of taps[13,14].	
buttonNone		Frequency of tapping outside the button [13,14].	
corXY		Correlation between X and Y position of tap on screen coordinates [13,14].	
mean	s	Mean value of the observations[4,13,14].	Intertap interval (TapInter), Leftdrift (DriftLeft), Right drift (DriftRight)
min	s	Minimum value of the observations[4,13,14].	
max	s	Maximum value of the observations[4,13,14].	
median	s	Median value of the observations[4,13,14].	
mad	s	Median absolute deviation[13,14].	
sd	s	Standard deviation[4,13,14].	
range	s	Range of the observations[4,13,14].	
iqr	s	Interquartile range [13,14].	
cv		Coefficient of variation [4,13,14].	
skew		Skewness[13,14].	
kur		Kurtosis[13,14].	
tkeo		Teager-Kaiser Energy Operator. Measures energy variation[4,13,14].	Intertap interval (TapInter)
dfa		Detrended Fluctuation Analysis. Measures changes in the signal[4,13,14].	
ar1		Coefficient of an autoregressive model at lag 1. Indicates associations between intertap intervals[4,13,14].	
ar2		Coefficient of an autoregressive model at lag 2. Indicates associations between intertap intervals[4,13,14].	
fatigue10		Increase in the mean intertap interval from the first 10% to the last 10% taps [4,13,14].	
fatigue25		Increase in the mean intertap interval from the first 25% to the last 25% taps[4,13,14].	
fatigue50		Increase in the mean intertap interval from the first 50% to the last 50 % taps[4,13,14].	

Table S5. Results from Mann-Whitney U test, Cohen's d and Median ICC for Gait task. Median ICC across different time points and different repetitions for PD and HC.

Feature Name	Mann-Whitney	|Cohens'd|	Median ICC			
	U test		Time Point		Repetition	
	P Value		HC	PD	HC	PD
frec_peak_LB_a	<0.001	0.3564	0.3213	0.2922	0.4335	0.3019
FreezeInd_z	<0.001	0.281	0.3706	0.2629	0.3807	0.3129

iqr_x	<0.001	0.2801	0.2242	0.2217	0.2521	0.2769
MSI	<0.001	0.2619	0.15	0.2869	0.24	0.231
numSteps	<0.001	0.2295	0.1417	0.2519	0.218	0.2265
median_acc	<0.001	0.225	0.2452	0.1278	0.1907	0.1501
PeakEnerg_LB_x	<0.001	0.2223	0.2338	0.2157	0.2458	0.2515
min_z	0.001	0.2205	0.2101	0.171	0.2324	0.2106
ApEn_pos_z	0.003	0.2127	0.0521	0.1015	0.0599	0.1263
Power_LB_x	0.001	0.2105	0.2141	0.2541	0.2726	0.3131
Power_FB_z	<0.001	0.2088	0.4055	0.2066	0.3187	0.317
ApEn_pos_a	0.006	0.2047	0.0239	0.1807	0.0461	0.1396
mean_a	0.029	0.1903	0.138	0.18	0.1776	0.2488
iqr_acc	0.003	0.1833	0.0641	0.2097	0.1114	0.2566
mean_acc	0.004	0.1818	-0.0834	0.0573	0.0267	-0.0259
rms_acc	0.012	0.1736	-0.0149	0.1103	0.1055	0.1264
rms_y	0.013	0.1707	0.1466	0.1984	0.2373	0.228
PeakEnerg_LB_a	<0.001	0.1674	0.3301	0.3331	0.3356	0.3769
RatioPower_y	0.029	0.1646	0.1631	0.2649	0.2408	0.2981
kur_pos_z	0.003	0.1628	0.1048	0.055	0.0521	0.0607
frec_peak_FB_vel_10	0.004	0.1609	0.0693	0.0929	0.0728	0.1018
iqr_z	0.005	0.1603	0.2659	0.2444	0.2908	0.3284
skew_z	0.003	0.1589	0.3317	0.1838	0.3627	0.2843
frec_peak_FB_a	0.001	0.1574	0.1287	0.1786	0.1707	0.1602
median_a	0.023	0.157	0.1741	0.262	0.2448	0.3192
FreezeInd_x	0.003	0.1568	0.2882	0.2689	0.3965	0.255
Power_LB_z	<0.001	0.1553	0.2284	0.2034	0.3033	0.2472
max_acc	0.022	0.1538	0.1236	0.1021	0.1352	0.133
zcr_x	0.002	0.1503	0.3311	0.3121	0.4098	0.3866
Power_LB_y	0.036	0.1424	0.2764	0.3183	0.3445	0.3732
Power_FB_acc	0.019	0.1416	0.1886	0.1934	0.1689	0.2459
kur_pos_a	0.034	0.1407	0.0457	0.0972	0.0852	0.1431
iqr_y	0.019	0.1357	0.2813	0.3066	0.3125	0.367
FreezeInd_a	0.039	0.1321	0.2859	0.2453	0.321	0.2287

Power_FB_vel	0.009	0.1301	0.0782	0.132	0.1196	0.1779
kur_x	0.029	0.1295	0.268	0.2131	0.3992	0.3336
frec_peak_LB_z	0.001	0.129	0.2095	0.3273	0.3098	0.3143
COEFCEPS20_pos_y	0.008	0.1284	0.0739	0.0609	0.0456	0.0069
COEFCEPS7_z	0.004	0.1243	-0.0383	0.057	-0.0183	0.038
PeakEnerg_FB_x	0.039	0.122	0.3281	0.1444	0.3445	0.2341
skew_vel	0.013	0.1213	0.1296	0.1193	0.0878	0.1066
cov_vel	0.013	0.1188	0.101	0.1169	0.0748	0.114
COEFCEPS1_pos_x	0.017	0.1173	-0.0056	-0.0537	0.0362	0.0325
iqr_pos_z	0.021	0.1137	0.0342	0.0343	0.0799	0.0666
cov_a	0.009	0.1129	0.4435	0.4726	0.4793	0.5028
FreezeInd_pos_a	0.022	0.1108	0.0767	0.0169	0.1282	0.0308
COEFCEPS9_z	0.04	0.1107	-0.007	-0.0297	-0.0125	-0.0148
PeakEnerg_LB_z	0.004	0.1076	0.2642	0.2013	0.2253	0.2781
COEFCEPS8_z	0.045	0.107	0.0796	0.0194	-0.003	-0.0192
COEFCEPS6_z	0.05	0.1062	-0.1046	0.0624	-0.0412	0.0281
frec_peak_LB_y	<0.001	0.1048	0.175	0.1178	0.2701	0.1949
ApEn_vel	0.015	0.1028	0.1951	0.1016	0.1224	0.069
ApEn_pos_x	0.015	0.0968	-0.0401	0.0387	-0.0158	0.0714
median_y	0.043	0.0926	0.1698	0.2399	0.2648	0.3101
COEFCEPS10_z	0.036	0.0922	-0.0703	-0.0739	0.0019	-0.0175
PeakEnerg_LB_pos_z	<0.001	0.0904	0.0393	0.0388	0.0533	0.0057
COEFCEPS1_pos_a	0.048	0.0882	0.0508	0.0852	0.0656	0.0644
zcr_pos_z	0.049	0.0814	0.0262	0.1287	0.0571	0.1752
RatioPower_pos_z	0.002	0.075	-0.0051	0.0112	0.0302	-0.0045
RatioPower_pos_a	0.038	0.0726	-0.0101	-0.0107	0.0188	-0.0098
Power_FB_pos_z	0.014	0.0721	-0.0136	0.008	0.0269	-0.0055
ApEn_pos_y	0.037	0.0656	0.088	0.121	0.0823	0.099
kur_vel	0.039	0.0587	0.2061	0.0857	0.0866	0.0433
cov_acc	0.048	0.037	-0.0013	-0.0008	-0.0096	-0.003

min_pos_x	0.021	0.0359	-0.0302	0.0342	-0.0335	0.0243
min_pos_y	0.023	0.0027	-0.0083	0.0136	-0.023	-0.0153

PD- Parkinson’s Disease, HC- Healthy Control, frec_peak- Frequency at the Peak of Energy, FreezeInd- Freeze Index, iqr-Interquartile Range, MSI- Mean Stride Interval, numSteps- Number of Steps, PeakEnerg- Peak of Energy, ApEn- Entropy, rms- Root Mean Square, RatioPower - Sum of the Power in the Freezing and Locomotor Band, skew- Skewness, min- Minimum Value, cov-Coefficient of Variation, zcr- Zero-Crossing Rate, kur-Kurtosis, COEFCEPS (1-20)- Mel Frequency Cepstral Coefficients, ar- Coefficient of a 1st Order Autoregressive Model, LB- Locomotor Band, FB- Freezing Band, vel- Velocity, acc- Acceleration Along Path, a- Accelerometer Average Signal, x- Accelerometer Mediolateral Signal, y- Accelerometer Vertical Signal, z- Accelerometer Anteroposterior Signal.

Table S6. Results from Mann-Whitney U test, Cohen’s d and Median ICC for features from Balance task. Median ICC across different time points and different repetitions for PD and HC. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	*Mann-Whitney U Test P Value*	|Cohens'd|	Median ICC			
			Time Point		Repetition	
			HC	PD	HC	PD
Power_MF_trem_z	<0.001	0.337	0.2808	0.2956	0.3016	0.2844
PeakEnerg_VHF_trem_x	0.028	0.3129	0.2521	0.3188	0.159	0.2677
PeakEnerg_VHF_trem_z	<0.001	0.2942	0.3225	0.2267	0.2318	0.205
RHL_trem_z	<0.001	0.2679	0.2891	0.5605	0.2936	0.1331
RHL_trem_a	0.004	0.2517	0.4071	0.3245	0.4463	0.294
Power_trem_y	<0.001	0.2441	0.0492	0.1998	0.1737	0.1535
F95_post_y	<0.001	0.2402	0.1742	0.1308	0.2351	0.1838
Power_trem_z	0.006	0.2317	0.2601	0.2376	0.2824	0.2763
median_trem_a	<0.001	0.2203	0.0973	0.1609	0.1475	0.1638
Power_LF_trem_z	<0.001	0.2186	0.2016	0.1727	0.2414	0.1817
CFREQ_post_z	0.008	0.2057	0.3047	0.2654	0.3369	0.2716
F95_post_x	0.002	0.1956	0.1938	0.1679	0.2122	0.1707
MFREQ_dist_x	<0.001	0.1915	0.0429	0.0269	0.0494	0.0648
iqr_post_y	<0.001	0.1873	0.1721	0.2481	0.1837	0.2804
mean_trem_a	<0.001	0.1851	0.1015	0.1805	0.1317	0.2164
kur_trem_x	<0.001	0.1799	0.0749	0.0185	0.0634	0.0222
F95_post_a	0.012	0.1776	0.1597	0.1905	0.2026	0.2202
ApEn_trem_x	<0.001	0.1757	0.0841	0.091	0.1253	0.0905

zcr_post_y	<0.001	0.1733	0.2208	0.2434	0.2412	0.1457
median_post_a	<0.001	0.1724	0.1195	0.0992	0.1907	0.1524
iqr_trem_x	0.002	0.1699	0.16	0.1447	0.1766	0.1522
FD_CC_dist_x_z	<0.001	0.1638	0.0852	-0.0044	0.0214	0.0247
F50_post_y	0.001	0.1631	0.1875	0.209	0.2339	0.2354
Power_LF_trem_x	0.01	0.1608	0.1362	0.1831	0.2339	0.1794
range_trem_y	<0.001	0.1585	0.1733	0.0855	0.2059	0.1177
MVELO_dist_x	0.006	0.1573	0.1408	0.1285	0.2058	0.1832
mean_post_y	0.007	0.1532	0.3668	0.2441	0.3952	0.2764
min_post_y	<0.001	0.1521	0.2216	0.2646	0.2126	0.2841
iqr_post_x	0.007	0.1449	0.1677	0.1459	0.1918	0.1951
kur_post_x	0.002	0.1391	0.0707	0.0247	0.0703	0.0539
rms_trem_a	0.004	0.1372	0.1666	0.152	0.1892	0.2266
ApEn_post_a	0.004	0.1361	0.0435	0.1382	0.1055	0.1037
skew_post_a	<0.001	0.1317	0.0291	0.1097	0.0483	0.0813
AREA_CC_dist_x_z	0.013	0.1306	0.1307	0.0011	0.0632	0.0322
FD_dist_x_z	0.003	0.1303	0.0781	-0.0472	0.019	0.0282
cov_trem_a	<0.001	0.1289	0.0988	0.032	0.1202	0.0325
TotalPower_post_x_z	0.011	0.1277	0.156	0.0493	0.1804	0.0726
MFREQ_dist_x_z	0.001	0.1261	0.0836	-0.0104	0.0415	0.0331
max_post_y	<0.001	0.1245	0.1192	0.2187	0.2363	0.2089
F50_post_x	<0.001	0.1243	0.2149	0.253	0.1823	0.29
cov_post_a	0.009	0.1199	0.0641	0.0677	0.0992	0.0634
range_trem_x	0.05	0.1195	0.1877	0.2015	0.2212	0.1329
ApEn_trem_y	0.001	0.1146	0.0886	0.0736	0.1512	0.1158
FD_CE_dist_x_z	0.007	0.1055	0.1519	0.0453	0.0895	0.0774
mean_post_z	0.031	0.0994	0.1935	0.1906	0.1498	0.1938
F50_post_a	0.003	0.0988	0.2095	0.2191	0.243	0.2912
Power_LF_trem_a	0.005	0.0982	0.217	0.0347	0.185	0.0634
max_post_z	0.027	0.0939	0.207	0.1773	0.1862	0.2413
rms_post_a	0.028	0.0925	0.2293	0.1188	0.2554	0.1798
F50_post_x_z	0.006	0.0882	0.2172	0.2453	0.2351	0.3015

FREQD_post_x	0.005	0.0848	0.1709	0.2444	0.1828	0.2444
TotalPower_post_z	0.014	0.0837	0.226	0.1239	0.2691	0.192
FREQD_post_x_z	0.019	0.0811	0.1243	0.1909	0.1926	0.1707
TotalPower_post_y	<0.001	0.0805	0.132	0.1737	0.216	0.2851
kur_post_a	<0.001	0.0717	0.0227	0.0732	0.0094	0.0578
max_trem_z	0.038	0.0709	0.0628	0.0578	0.0982	0.1206
jerk_post_y	<0.001	0.0421	0.0474	0.0484	0.181	0.1174
kur_trem_a	0.012	0.0349	0.0081	0.0034	0.0014	-0.0098
kur_trem_y	0.022	0.0233	0.0305	-0.0257	0.0612	-0.0137

PD- Parkinson's Disease, HC- Healthy Control, PeakEnergy - Peak of energy, TotalPower- Energy between 15-3.5 Hz, rms- Root Mean Square, F50- Frequency Containing 50% of Total Power, F95- Frequency containing 95% of the total power, FRQD- Frequency of Dispersion of the Power Spectrum, MFREQ- Mean Frequency, iqr- Interquartile Range, kur- Kurtosis, zcr- Zero-Crossing Rate, ApEn- Entropy, skew- Skewness, jerk- Average jerk, MVELO- Mean velocity, FD- Fractal Dimension, FD_CE- Fractal Dimension based on the 95% Confidence Ellipse Area, min- Minimum Value, CFREQ- Centroidal Frequency, RHL- Ratio Between Power in High Frequency and Low Frequency, dist- Distance, 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.

Table S7. Results from Mann-Whitney U test, Cohen's d and Median ICC for features from Voice task. Median ICC across different time points and different repetitions for PD and HC. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	Mann-Whitney	|Cohens'd|	Median ICC			
	U Test		Time Point		Repetition	
	P Value		HC	PD	HC	PD
mean_gqc	<0.001	0.5049	0.6737	0.7093	0.6797	0.7476
p5_gqc	<0.001	0.412	0.3308	0.37	0.3169	0.4305
c_mean_MFCC1	<0.001	0.4028	0.5158	0.4304	0.5874	0.5071
p95_gqc	<0.001	0.354	0.6699	0.6128	0.6779	0.6494
std_tkeo	<0.001	0.3033	0.4022	0.3972	0.4577	0.4393
p95_tkeo	<0.001	0.2919	0.318	0.3387	0.3596	0.397
c_std_d11	<0.001	0.278	0.3747	0.3887	0.4209	0.4159
hnr_std	<0.001	0.2738	0.2193	0.2973	0.2553	0.3413
c_std_d12	<0.001	0.2717	0.3335	0.4011	0.4497	0.4253
p75_tkeo	<0.001	0.2679	0.4344	0.2648	0.4264	0.3476

c_std_d13	<0.001	0.2634	0.3766	0.3868	0.4272	0.4303
c_std_d8	<0.001	0.2491	0.3019	0.345	0.3748	0.3711
c_std_d9	<0.001	0.2445	0.3108	0.3664	0.3589	0.3817
c_mean_MFCC10	<0.001	0.2403	0.5648	0.6197	0.6255	0.6454
c_std_d10	<0.001	0.2401	0.3423	0.389	0.4058	0.4132
c_std_d7	<0.001	0.2309	0.3589	0.3308	0.3779	0.3758
c_std_dd11	<0.001	0.2282	0.4119	0.4281	0.4627	0.4498
c_std_d5	<0.001	0.2196	0.3335	0.2868	0.3318	0.3435
c_std_d14	<0.001	0.2193	0.3474	0.3912	0.3987	0.4291
c_std_d6	<0.001	0.2186	0.3699	0.3492	0.3717	0.3718
c_mean_MFCC12	<0.001	0.2154	0.4902	0.433	0.5362	0.4977
c_std_dd5	<0.001	0.2048	0.373	0.369	0.4207	0.4006
c_std_d3	<0.001	0.202	0.2891	0.3135	0.2862	0.3391
c_std_d4	<0.001	0.2008	0.3006	0.3027	0.3129	0.3435
c_std_dd10	<0.001	0.1976	0.4096	0.4436	0.4759	0.4647
c_std_dd9	<0.001	0.1942	0.3812	0.4232	0.4336	0.4327
c_std_dd8	<0.001	0.1932	0.3741	0.4008	0.454	0.4177
c_std_dd12	<0.001	0.1923	0.4006	0.446	0.4961	0.4601
c_std_MFCC1	0.026	0.1883	0.2528	0.251	0.2488	0.3304
c_std_MFCC5	0.001	0.1853	0.2922	0.2074	0.3001	0.2461
c_std_dd7	<0.001	0.1842	0.3847	0.3989	0.4603	0.4339
c_std_dd13	<0.001	0.1821	0.3979	0.4326	0.4806	0.4675
c_std_dd6	<0.001	0.182	0.3975	0.3964	0.4414	0.4133
DFA	<0.001	0.1792	0.0577	0.0833	0.0365	0.1416
c_mean_0th	0.002	0.1675	0.1974	0.2609	0.2692	0.2913
c_mean_d13	<0.001	0.1673	0.1084	0.1388	0.1036	0.127
fm	0.001	0.1648	0.3646	0.3997	0.3755	0.3536
c_std_dd4	0.002	0.1631	0.3667	0.3165	0.3905	0.3851
c_mean_MFCC7	0.001	0.1615	0.5329	0.5577	0.6043	0.5592
c_mean_d4	0.002	0.1597	0.0228	0.1646	0.0771	0.1991
shdb	0.03	0.1552	0.3822	0.4779	0.5244	0.523
c_std_dd3	0.001	0.1495	0.3525	0.3232	0.3634	0.3412
ApEn_f0	<0.001	0.148	0.0964	0.145	0.1605	0.2034

c_std_dd14	0.006	0.1447	0.3701	0.4477	0.4625	0.4762
c_mean_d3	0.001	0.1344	0.1441	0.1025	0.1508	0.1771
c_std_MFCC11	0.016	0.1323	0.231	0.2937	0.3196	0.3401
c_std_d1	0.031	0.1292	0.2985	0.3704	0.2786	0.3606
c_std_MFCC6	0.005	0.1261	0.2722	0.2996	0.3064	0.333
c_mean_MFCC5	0.008	0.1228	0.5375	0.5294	0.5745	0.5899
c_mean_MFCC9	0.004	0.1209	0.4986	0.4885	0.5747	0.5466
c_std_MFCC9	0.014	0.1172	0.2365	0.2543	0.285	0.2889
rpde	0.002	0.1161	0.3277	0.3635	0.3636	0.3956
c_std_MFCC7	0.041	0.1087	0.2765	0.3007	0.3085	0.3234
c_mean_MFCC8	0.048	0.0999	0.4376	0.5004	0.5097	0.524
c_std_MFCC8	0.028	0.0934	0.1816	0.3283	0.2736	0.3469
c_mean_MFCC6	0.019	0.081	0.5523	0.5853	0.587	0.5872
c_mean_d2	0.002	0.0781	0.1124	0.1832	0.1251	0.2258
c_mean_d6	0.039	0.0725	0.1365	0.1728	0.1582	0.1926
c_mean_MFCC3	0.02	0.0706	0.4717	0.5061	0.5486	0.5317
c_mean_d1	0.005	0.0705	0.1218	0.1949	0.1002	0.2329

PD- Parkinson's Disease, HC- Healthy Control, c_mean- Mean of the MFCCs Coefficients, log-Energy of the Signal and the First and Second Derivatives of the MFCCs, MFCC- Mel Frequency Cepstral Coefficients, c_std- Standard Deviation of the MFCCs Coefficients, gqc- Glottis Quotient Close, fm- Frequency Modulation, std - Standard Deviation, tkeo- Teager Kaiser Energy Operator, p5- 5th percentile, p75- 75th Percentile, p95- 95th Percentile, shbd- Shimmer, hnr- Harmonic to Noise Ratio, d- Delta, d-d- Delta-Delta, DFA- Detrended Fluctuation Analysis, f0- Fundamental Frequency, ApEn- Pitch Period Entropy, rpde- Recurrence Period Density Entropy, T- Period.

Table S8. Results from Mann-Whitney U test, Cohen's d and Median ICC for features from Tapping task. Median ICC across different time points and different repetitions for PD and HC. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	Mann-Whitney U test *P Value*	|Cohens'd |	Median ICC			
			Time Point		Repetition	
			HC	PD	HC	PD
numberTaps	<0.001	1.179	0.6896	0.6411	0.6895	0.7004
max_TapInter	<0.001	0.6177	0.1792	0.2758	0.1838	0.2909
range_TapInter	<0.001	0.5821	0.1297	0.2741	0.1628	0.2809

ar2_TapInter	<0.001	0.5322	0.3102	0.2777	0.3286	0.3243
ar1_TapInter	<0.001	0.5282	0.3039	0.2938	0.3193	0.3217
sd_TapInter	<0.001	0.5154	0.2779	0.2258	0.3023	0.3459
buttonNoneFreq	<0.001	0.4455	0.3076	0.3658	0.3453	0.4613
mad_TapInter	<0.001	0.3701	0.3294	0.305	0.3488	0.3751
median_DriftRight	<0.001	0.3542	0.6796	0.5393	0.4578	0.5345
mad_DriftRight	<0.001	0.3468	0.4492	0.4259	0.466	0.4822
median_DriftLeft	<0.001	0.2912	0.51	0.5269	0.5202	0.5292
min_TapInter	<0.001	0.2857	0.2751	0.3967	0.4364	0.4605
sd_DriftRight	<0.001	0.2842	0.4088	0.2833	0.4066	0.3227
iqr_TapInter	<0.001	0.2785	0.4353	0.3046	0.4745	0.3506
mad_DriftLeft	<0.001	0.2725	0.4603	0.4723	0.4801	0.4591
sd_DriftLeft	<0.001	0.2694	0.3654	0.4028	0.3932	0.2981
skew_TapInter	<0.001	0.2315	0.1461	0.1855	0.1614	0.2029
skew_DriftLeft	<0.001	0.208	0.0637	0.1046	0.092	0.1149
kur_DriftLeft	<0.001	0.1913	0.0299	0.063	0.0924	0.0827
kur_DriftRight	<0.001	0.1842	0.1257	0.0258	0.1567	0.0796
kur_TapInter	<0.001	0.1695	0.092	0.1072	0.1043	0.1117
cv_TapInter	<0.001	0.1586	0.4259	0.4212	0.4716	0.4547
corXY	0.001	0.114	0.624	0.5737	0.5804	0.5324
tkeo_TapInter	<0.001	0.0905	0.0437	0.0275	0.0439	0.0988
cv_DriftLeft	0.014	0.0113	0.0981	0.135	0.1078	0.162

PD- Parkinson's Disease, HC- Healthy Control, 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 Absolute Deviation, min- Minimum, max- Maximum, skew- Skewness, kur- Kurtosis, teko- Teager-Kaiser Energy Operator, cv- Coefficient, Sd- Standard Deviation, ar (1-2)- Coefficient of an Autoregressive Model at Lag (1-2).

Table S9. Results from an analysis of ANOVA for repeated measurements on the most reliable features in PD and HC selected from different repetition.

Features	Diagnosis (PD, HC)		Repetition		Diagnosis × Repetition		Medication
	F	*P*	F	*P*	F	*P*	n=188
Gait Task							

frec_peak_LB_a	23.202	<0.001	0.767	0.547	0.276	0.894	0.358
FreezeInd_z	3.877	0.049	0.665	0.616	1.012	0.4	0.422
Power_FB_z	0.144	0.704	10.235	<0.001	4.719	<0.001	0.08
PeakEnerg_LB_a	10.268	0.001	9.974	<0.001	3.02	0.017	0.376
iqr_z	0.109	0.742	6.849	<0.001	1.102	0.354	0.806
skew_z	9.226	0.002	2.749	0.027	0.267	0.899	0.192
median_a	1.003	0.317	13.362	<0.001	2.814	0.024	0.058
FreezeInd_x	2.681	0.102	0.954	0.432	0.406	0.804	0.151
zcr_x	4.987	0.026	4.387	0.002	0.827	0.508	0.58
Power_LB_y	0.256	0.613	2.985	0.018	0.792	0.53	0.156
iqr_y	0	1.0	2.28	0.059	1.223	0.299	0.01
kur_x	2.438	0.119	0.418	0.796	1.969	0.097	0.375
frec_peak_LB_z	1.013	0.315	2.198	0.067	0.705	0.589	0.449
PeakEnerg_FB_x	9.421	0.002	12.071	<0.001	0.979	0.418	0.028
cov_a	1.463	0.227	1.341	0.252	1.493	0.202	0.844
			Balance Task				
Features	F	*P*	F	*P*	F	*P*	n=189
Power_MF_trem_z	29.608	<0.001	1.096	0.357	0.549	0.7	0.192
RHL_trem_z	19.372	<0.001	1.848	0.117	1.712	0.145	0.349
RHL_trem_a	29.005	<0.001	1.442	0.217	0.254	0.907	0.108
Power_trem_z	13.015	<0.001	2.885	0.021	0.526	0.717	0.143
Power_LF_trem_z	15.785	<0.001	5.59	<0.001	0.901	0.462	0.614
CFREQ_post_z	9.483	0.002	1.532	0.19	0.772	0.544	0.066
iqr_post_y	33.914	<0.001	10.249	<0.001	0.609	0.656	0.245
mean_post_y	5.178	0.023	11.917	<0.001	0.328	0.859	<0.001
min_post_y	20.291	<0.001	16.579	<0.001	0.882	0.474	0.555
F50_post_x	0.975	0.324	0.631	0.641	1.372	0.241	0.285
F50_post_a	0.048	0.827	0.607	0.657	1.1	0.355	0.561
rms_post_a	16.826	<0.001	5.588	<0.001	0.869	0.482	0.603
F50_post_x_z	0.006	0.938	0.157	0.96	1.421	0.224	0.106

TotalPower_post_z	2.378	0.124	3.339	0.01	0.404	0.806	0.684
TotalPower_post_y	14.926	<0.001	7.7	<0.001	0.091	0.985	0.958
Voice Task							
Features	F	*P*	F	*P*	F	*P*	n=280
mean_gqc	92.952	<0.001	3.284	0.011	2.551	0.037	0.108
c_mean_MFCC1	38.813	<0.001	24.89	<0.001	1.836	0.119	0.611
p95_gqc	59.181	<0.001	1.531	0.19	1.706	0.146	0.016
c_mean_MFCC10	22.916	<0.001	0.796	0.528	0.586	0.673	0.257
c_mean_MFCC12	10.811	0.001	2.232	0.063	1.304	0.266	0.012
c_mean_MFCC7	13.614	<0.001	1.009	0.401	1.73	0.14	0.182
shdb	7.787	0.005	1.985	0.094	0.538	0.708	0.375
c_mean_MFCC5	7.142	0.008	2.813	0.024	0.198	0.939	0.084
c_mean_MFCC9	1.264	0.261	1.554	0.184	0.205	0.936	0.186
c_mean_MFCC8	1.373	0.242	1.375	0.24	0.552	0.698	0.676
c_mean_MFCC6	0.936	0.334	2.035	0.087	1.011	0.4	0.982
c_mean_MFCC3	1.886	0.17	7.87	<0.001	6.754	<0.001	0.249
Tapping Task							
	F	*P*	F	*P*	F	*P*	n=338
numberTaps	539.151	<0.001	123.309	<0.001	4.444	0.001	<0.001
buttonNoneFreq	50.866	<0.001	4.93	<0.001	2.344	0.052	0.929
mad_TapInter	87.645	<0.001	0.721	0.578	4.172	0.002	0.593
median_DriftRight	8.987	0.003	9.747	<0.001	6.351	<0.001	0.88
mad_DriftRight	8.9	0.003	10.605	<0.001	6.353	<0.001	0.846
median_DriftLeft	4.491	0.034	3.107	0.015	2.769	0.026	0.683
min_TapInter	44.616	<0.001	11.437	<0.001	0.388	0.818	0.226
sd_DriftRight	4.338	0.037	6.689	<0.001	3.516	0.007	0.279
iqr_TapInter	53.591	<0.001	1.054	0.378	2.553	0.037	0.319
mad_DriftLeft	2.594	0.108	11.145	<0.001	6.06	<0.001	0.548
cv_TapInter	8.054	0.005	6.815	<0.001	1.702	0.146	0.232
corXY	5.935	0.015	21.815	<0.001	2.541	0.038	0.874

HC- Healthy Controls, PD- Parkinson’s Disease, M- Male, F- Female, S.D- Standard Deviation, 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, 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, shbd- Shimmer, gqc- Glottis Quotient Close, p95- 95th Percentile, 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 Absolute Deviation, min- Minimum, cv- Coefficient, Sd- Standard Deviation.

Table S10. Results from an analysis of ANOVA for repeated measurements on the features from Gait task. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	Medication		diagnosis (PD, HC)		Repetition		Diagnosis × Repetition	
	F	*P*	F	*P*	F	*P*	F	*P*
frec_peak_LB_a	1.03	0.358	23.202	<0.001	0.767	0.547	0.276	0.894
FreezeInd_z	0.864	0.422	3.877	0.049	0.665	0.616	1.012	0.4
iqr_x	0.348	0.706	1.502	0.221	12.892	<0.001	1.77	0.132
MSI	0.637	0.529	14.831	<0.001	0.61	0.655	0.113	0.978
numSteps	0.696	0.499	17.274	<0.001	1.038	0.386	0.067	0.992
median_acc	0.074	0.928	9.673	0.002	2.833	0.023	1.262	0.283
PeakEnerg_LB_x	0.104	0.901	0.003	0.954	12.381	<0.001	1.703	0.147
min_z	0.564	0.57	5.477	0.02	2.724	0.028	3.282	0.011
ApEn_pos_z	0.249	0.779	13.714	<0.001	0.701	0.591	1.472	0.208
Power_LB_x	1.85	0.159	0.215	0.643	8.244	<0.001	1.845	0.118
Power_FB_z	2.549	0.08	0.144	0.704	10.235	<0.001	4.719	<0.001
ApEn_pos_a	1.239	0.291	2.417	0.121	2.913	0.02	1.005	0.404
mean_a	3.781	0.024	1.053	0.305	14.267	<0.001	4.027	0.003
iqr_acc	3.521	0.031	0.014	0.904	10.881	<0.001	2.116	0.076
mean_acc	0.411	0.663	9.221	0.002	1.353	0.248	0.227	0.923

rms_acc	3.682	0.026	0.088	0.767	8.63	<0.001	4.71	<0.001
rms_y	3.048	0.049	0.433	0.511	6.438	<0.001	4.419	0.001
PeakEnerg_LB_a	0.982	0.376	10.268	0.001	9.974	<0.001	3.02	0.017
RatioPower_y	2.183	0.114	1.898	0.169	8.466	<0.001	3.447	0.008
kur_pos_z	0.724	0.486	10.965	<0.001	0.923	0.45	2.679	0.03
frec_peak_FB_vel_10	0.34	0.712	7.353	0.007	3.544	0.007	1.017	0.397
iqr_z	0.215	0.806	0.109	0.742	6.849	<0.001	1.102	0.354
skew_z	1.658	0.192	9.226	0.002	2.749	0.027	0.267	0.899
frec_peak_FB_a	1.067	0.345	10.735	0.001	1.114	0.348	0.357	0.839
median_a	2.875	0.058	1.003	0.317	13.362	<0.001	2.814	0.024
FreezeInd_x	1.899	0.151	2.681	0.102	0.954	0.432	0.406	0.804
Power_LB_z	0.264	0.768	9.674	0.002	11.043	<0.001	1.32	0.26
max_acc	0.167	0.846	2.783	0.096	1.306	0.265	3.298	0.011
zcr_x	0.546	0.58	4.987	0.026	4.387	0.002	0.827	0.508
Power_LB_y	1.865	0.156	0.256	0.613	2.985	0.018	0.792	0.53
Power_FB_acc	0.771	0.463	1.372	0.242	9.907	<0.001	5.476	<0.001
kur_pos_a	0.934	0.394	0.885	0.347	4.982	<0.001	2.408	0.047
iqr_y	4.641	0.01	0	1.0	2.28	0.059	1.223	0.299
FreezeInd_a	0.105	0.9	2.124	0.146	1.102	0.354	0.584	0.674
Power_FB_vel	0.797	0.452	0.585	0.445	1.37	0.242	1.374	0.24
kur_x	0.984	0.375	2.438	0.119	0.418	0.796	1.969	0.097
frec_peak_LB_z	0.802	0.449	1.013	0.315	2.198	0.067	0.705	0.589
COEFCEPS20_pos_y	0.227	0.797	0	1.0	3.391	0.009	0.51	0.729
COEFCEPS7_z	0.11	0.896	1.377	0.241	0.548	0.7	2.725	0.028
PeakEnerg_FB_x	3.604	0.028	9.421	0.002	12.071	<0.001	0.979	0.418
skew_vel	0.582	0.559	20.913	<0.001	3.57	0.007	2.99	0.018
cov_vel	0.67	0.512	12.713	<0.001	2.013	0.09	2.824	0.024
COEFCEPS1_pos_x	0.171	0.843	13.528	<0.001	3.217	0.012	0.333	0.856
iqr_pos_z	0.65	0.523	3.476	0.063	6.916	<0.001	0.898	0.464
cov_a	0.17	0.844	1.463	0.227	1.341	0.252	1.493	0.202
FreezeInd_pos_a	0.355	0.702	0	1.0	2.16	0.071	1.584	0.176
COEFCEPS9_z	0.153	0.858	0.365	0.546	0.749	0.559	1.771	0.132
PeakEnerg_LB_z	0.13	0.878	6.525	0.011	11.047	<0.001	0.405	0.805

COEFCEPS8_z	0.086	0.918	0	1.0	0.996	0.408	2.3	0.057
COEFCEPS6_z	0.675	0.51	0.54	0.463	0.463	0.763	1.135	0.338
frec_peak_LB_y	0.396	0.674	1.887	0.17	3.161	0.013	0.116	0.977
ApEn_vel	0.54	0.583	18.003	<0.001	2.215	0.065	2.301	0.056
ApEn_pos_x	0.846	0.43	11.025	<0.001	1.571	0.179	2.071	0.082
median_y	0.556	0.574	0.268	0.605	2.994	0.018	0.024	0.999
COEFCEPS10_z	0.231	0.794	0.108	0.742	0.859	0.488	1.824	0.122
PeakEnerg_LB_pos_z	0.381	0.683	4.483	0.035	3.138	0.014	0.575	0.681
COEFCEPS1_pos_a	0.496	0.609	3.953	0.047	1.476	0.207	0.818	0.514
zcr_pos_z	0.546	0.58	0	1.0	0	1.0	5.182	<0.001
RatioPower_pos_z	0.483	0.617	2.316	0.129	3.968	0.003	0.359	0.838
RatioPower_pos_a	1.356	0.259	1.389	0.239	7.854	<0.001	0.326	0.861
Power_FB_pos_z	0.491	0.613	1.871	0.172	4.19	0.002	0.321	0.864
ApEn_pos_y	0.214	0.807	4.055	0.045	2.091	0.079	0.558	0.693
kur_vel	0.449	0.639	14.725	<0.001	1.367	0.243	2.074	0.082
cov_acc	0.433	0.649	0.175	0.676	1.029	0.391	0.864	0.485
min_pos_x	4.193	0.016	0.692	0.406	2.735	0.027	0.908	0.458
min_pos_y	2.354	0.096	0.069	0.793	6.382	<0.001	1.057	0.377

PD- Parkinson’s Disease, HC- Healthy Control, frec_peak- Frequency at the Peak of Energy, FreezeInd- Freeze Index, iqr-Interquartile Range, MSI- Mean Stride Interval, numSteps- Number of Steps, PeakEnerg- Peak of Energy, ApEn- Entropy, rms- Root Mean Square, RatioPower - Sum of the Power in the Freezing and Locomotor Band, skew- Skewness, min- Minimum Value, cov-Coefficient of Variation, zcr- Zero-Crossing Rate, kur-Kurtosis, COEFCEPS (1-20)- Mel Frequency Cepstral Coefficients, ar- Coefficient of a 1st Order Autoregressive Model, LB- Locomotor Band, FB- Freezing Band, vel- Velocity, acc- Acceleration Along Path, a- Accelerometer Average Signal, x- Accelerometer Mediolateral Signal, y- Accelerometer Vertical Signal, z- Accelerometer Anteroposterior Signal.

Table S11. Results from an analysis of ANOVA for repeated measurements on the features from Balance task. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	Medication		Diagnosis (PD, HC)		Repetition		diagnosis × Repetition	
	F	*P*	F	*P*	F	*P*	F	*P*
Power_MF_trem_z	1.657	0.192	29.608	<0.001	1.096	0.357	0.549	0.7
PeakEnerg_VHF_trem_x	0.643	0.526	31.861	<0.001	0.165	0.956	0.345	0.848

PeakEnerg_VHF_trem_z	0.934	0.394	26.313	<0.001	0.927	0.447	0.775	0.542
RHL_trem_z	1.055	0.349	19.372	<0.001	1.848	0.117	1.712	0.145
RHL_trem_a	2.235	0.108	29.005	<0.001	1.442	0.217	0.254	0.907
Power_trem_y	0.618	0.54	25.883	<0.001	5.182	<0.001	0.295	0.882
F95_post_y	0.337	0.714	17.64	<0.001	6.648	<0.001	0.241	0.915
Power_trem_z	1.956	0.143	13.015	<0.001	2.885	0.021	0.526	0.717
median_trem_a	1.993	0.138	41.154	<0.001	0.738	0.566	0.471	0.757
Power_LF_trem_z	0.488	0.614	15.785	<0.001	5.59	<0.001	0.901	0.462
CFREQ_post_z	2.744	0.066	9.483	0.002	1.532	0.19	0.772	0.544
F95_post_x	1.179	0.309	3.449	0.064	3.528	0.007	0.317	0.867
MFREQ_dist_x	0.153	0.858	19.47	<0.001	0	1.0	0.341	0.851
iqr_post_y	1.41	0.245	33.914	<0.001	10.249	<0.001	0.609	0.656
mean_trem_a	1.037	0.355	34.677	<0.001	1.947	0.1	0.934	0.443
kur_trem_x	1.231	0.293	3.437	0.064	1.804	0.125	0.427	0.789
F95_post_a	2.29	0.103	4.361	0.037	0.532	0.712	0.198	0.939
ApEn_trem_x	0.806	0.447	9.629	0.002	4.215	0.002	0.137	0.969
zcr_post_y	0.05	0.952	16.755	<0.001	0	1.0	-1.081	1
median_post_a	1.58	0.207	30.751	<0.001	3.102	0.015	0.698	0.593
iqr_trem_x	3.391	0.035	7.537	0.006	2.05	0.085	0.343	0.849
FD_CC_dist_x_z	1.485	0.228	17.237	<0.001	0.948	0.435	0.819	0.513
F50_post_y	0.611	0.543	13.841	<0.001	4.862	<0.001	2.136	0.074
Power_LF_trem_x	0.317	0.729	3.233	0.073	3.255	0.011	1.005	0.404
range_trem_y	0.621	0.538	16.464	<0.001	12.134	<0.001	0.764	0.548
MVELO_dist_x	5.03	0.007	11.829	<0.001	12.194	<0.001	0.455	0.769
mean_post_y	4.806	0.009	5.178	0.023	11.917	<0.001	0.328	0.859
min_post_y	0.59	0.555	20.291	<0.001	16.579	<0.001	0.882	0.474
iqr_post_x	6.234	0.002	6.785	0.009	8.167	<0.001	0.463	0.763
kur_post_x	1.217	0.297	6.657	0.01	2.837	0.023	0.282	0.89
rms_trem_a	0.253	0.777	22.949	<0.001	2.598	0.035	0.798	0.527
ApEn_post_a	0.817	0.442	5.012	0.026	6.458	<0.001	0.616	0.651
skew_post_a	0.981	0.376	16.017	<0.001	3.468	0.008	0.685	0.602

AREA_CC_dist_x_z	0.229	0.796	21.152	<0.001	2.189	0.068	1.096	0.357
FD_dist_x_z	2.931	0.055	7.317	0.007	1.018	0.396	1.194	0.311
cov_trem_a	2.753	0.065	14.523	<0.001	1.629	0.164	0.562	0.69
TotalPower_post_x_z	0.188	0.829	4.258	0.039	1.226	0.298	0.496	0.738
MFREQ_dist_x_z	0.474	0.623	14.312	<0.001	0	1.0	-0.011	1
max_post_y	2.302	0.102	14.466	<0.001	13.553	<0.001	0.769	0.545
F50_post_x	1.259	0.285	0.975	0.324	0.631	0.641	1.372	0.241
cov_post_a	2.224	0.11	12.355	<0.001	2.121	0.076	0.056	0.994
range_trem_x	0.251	0.778	1.748	0.187	2.225	0.064	0.5	0.735
ApEn_trem_y	0.825	0.439	10.51	0.001	3.458	0.008	0.994	0.409
FD_CE_dist_x_z	2.346	0.097	0.523	0.47	1.258	0.284	0.383	0.821
mean_post_z	1.368	0.256	1.085	0.298	1.66	0.156	1.065	0.372
F50_post_a	0.579	0.561	0.048	0.827	0.607	0.657	1.1	0.355
Power_LF_trem_a	1.381	0.253	9.566	0.002	1.45	0.215	1.072	0.369
max_post_z	3.408	0.034	0.884	0.347	6.177	<0.001	0.531	0.713
rms_post_a	0.507	0.603	16.826	<0.001	5.588	<0.001	0.869	0.482
F50_post_x_z	2.256	0.106	0.006	0.938	0.157	0.96	1.421	0.224
FREQD_post_x	1.612	0.201	2.621	0.106	0.571	0.684	1.002	0.405
TotalPower_post_z	0.38	0.684	2.378	0.124	3.339	0.01	0.404	0.806
FREQD_post_x_z	0.951	0.387	0.258	0.612	0.194	0.942	0.084	0.987
TotalPower_post_y	0.043	0.958	14.926	<0.001	7.7	<0.001	0.091	0.985
kur_post_a	1.7	0.184	11.701	<0.001	2.441	0.045	2.134	0.074
max_trem_z	1.652	0.193	1.857	0.174	2.054	0.084	0.54	0.706
jerk_post_y	0.375	0.688	7.077	0.008	2.766	0.026	0.632	0.64
kur_trem_a	0.071	0.932	0	1.0	2.942	0.019	0.856	0.49
kur_trem_y	0.054	0.947	0.406	0.524	2.859	0.022	1.539	0.188

PD- Parkinson's Disease, HC- Healthy Control, PeakEnergy - Peak of energy, TotalPower- Energy between 15-3.5 Hz, rms- Root Mean Square, F50- Frequency Containing 50% of Total Power, F95- Frequency containing 95% of the total power, FRQD- Frequency of Dispersion of the Power Spectrum, MFREQ- Mean Frequency, iqr- Interquartile Range, kur- Kurtosis, zcr- Zero-Crossing Rate, ApEn- Entropy, skew- Skewness, jerk- Average jerk, MVELO- Mean velocity, FD- Fractal Dimension, FD_CE- Fractal Dimension based on the 95% Confidence Ellipse Area, min-

Minimum Value, CFREQ- Centroidal Frequency, RHL- Ratio Between Power in High Frequency and Low Frequency, dist- Distance, 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.

Table S12. Results from an analysis of ANOVA for repeated measurements on the features from Voice task. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	Medication		Diagnosis (PD, HC)		Repetition		Diagnosis × Repetition	
	F	P	F	P	F	P	F	P
mean_gqc	2.239	0.108	92.952	<0.001	3.284	0.011	2.551	0.037
p5_gqc	0.114	0.892	49.446	<0.001	1.938	0.101	0.545	0.703
c_mean_MFCC1	0.492	0.611	38.813	<0.001	24.89	<0.001	1.836	0.119
p95_gqc	4.158	0.016	59.181	<0.001	1.531	0.19	1.706	0.146
std_tkeo	0.176	0.838	2.341	0.126	3.392	0.009	1.89	0.109
p95_tkeo	0.543	0.581	4.843	0.028	1.932	0.102	0.874	0.478
c_std_d11	1.651	0.193	3.892	0.049	0	1.0	0.786	0.534
hnr_std	0.836	0.434	13.227	<0.001	18.399	<0.001	1.165	0.324
c_std_d12	1.265	0.283	4.053	0.044	0	1.0	4.556	0.001
p75_tkeo	3.092	0.046	1.517	0.218	2.435	0.045	2.042	0.086
c_std_d13	2.121	0.121	4.497	0.034	0	1.0	-2.873	1.0
c_std_d8	0.72	0.487	2.962	0.086	0	1.0	1.201	0.308
c_std_d9	1.669	0.189	3.243	0.072	0	1.0	0.978	0.418
c_mean_MFCC10	1.361	0.257	22.916	<0.001	0.796	0.528	0.586	0.673
c_std_d10	2.649	0.072	0	1.0	0	1.0	2.912	0.02
c_std_d7	1.583	0.206	2.291	0.13	0	1.0	2.004	0.091
c_std_dd11	2.3	0.101	0	1.0	0	1.0	6.108	<0.001
c_std_d5	1.988	0.138	8.873	0.003	6.115	<0.001	0.912	0.456
c_std_d14	1.869	0.155	0	1.0	0	1.0	3.318	0.01
c_std_d6	0.49	0.613	5.343	0.021	0	1.0	0.173	0.952
c_mean_MFCC12	4.491	0.012	10.811	0.001	2.232	0.063	1.304	0.266
c_std_dd5	2.368	0.095	0	1.0	0	1.0	6.746	<0.001
c_std_d3	2.208	0.111	12.446	<0.001	2.773	0.026	2.114	0.076

c_std_d4	1.978	0.139	6.739	0.01	3.227	0.012	0.614	0.653
c_std_dd10	3.345	0.036	0	1.0	0	1.0	6.864	<0.001
c_std_dd9	2.126	0.12	0	1.0	0	1.0	6.709	<0.001
c_std_dd8	1.622	0.198	0	1.0	0	1.0	6.034	<0.001
c_std_dd12	1.96	0.142	0	1.0	0	1.0	6.84	<0.001
c_std_MFCC1	0.931	0.395	13.851	<0.001	16.376	<0.001	1.183	0.316
c_std_MFCC5	2.659	0.071	1.814	0.178	4.874	<0.001	1.613	0.168
c_std_dd7	3.237	0.04	0	1.0	0	1.0	7.599	<0.001
c_std_dd13	3.163	0.043	0	1.0	0	1.0	6.79	<0.001
c_std_dd6	1.27	0.282	0	1.0	0	1.0	5.008	<0.001
DFA	0.298	0.742	10.114	0.002	1.763	0.133	1.062	0.374
c_mean_0th	0.815	0.443	26.062	<0.001	43.877	<0.001	1.754	0.135
c_mean_d13	0.43	0.651	0	1.0	0	1.0	3.693	0.005
fm	2.782	0.063	0.054	0.816	4.34	0.002	1.546	0.186
c_std_dd4	3.189	0.042	0	1.0	0	1.0	0.546	0.702
c_mean_MFCC7	1.71	0.182	13.614	<0.001	1.009	0.401	1.73	0.14
c_mean_d4	1.792	0.168	0	1.0	0	1.0	9.902	<0.001
shdb	0.982	0.375	7.787	0.005	1.985	0.094	0.538	0.708
c_std_dd3	4	0.019	0	1.0	0	1.0	4.832	<0.001
ApEn_f0	0.552	0.576	4.946	0.026	5.313	<0.001	1.242	0.291
c_std_dd14	2.764	0.064	0	1.0	0	1.0	10.367	<0.001
c_mean_d3	0.123	0.884	0	1.0	0	1.0	9.723	<0.001
c_std_MFCC11	0.292	0.747	0.139	0.709	3.71	0.005	0.946	0.436
c_std_d1	0.414	0.661	1.638	0.201	6.838	<0.001	1.692	0.149
c_std_MFCC6	0.936	0.393	1.571	0.21	5.619	<0.001	1.651	0.159
c_mean_MFCC5	2.483	0.084	7.142	0.008	2.813	0.024	0.198	0.939
c_mean_MFCC9	1.686	0.186	1.264	0.261	1.554	0.184	0.205	0.936
c_std_MFCC9	0.865	0.422	0.041	0.84	4.202	0.002	0.825	0.509
rpde	0.274	0.761	4.674	0.031	23.492	<0.001	5.175	<0.001
c_std_MFCC7	1.206	0.3	0.45	0.502	4.679	<0.001	1.148	0.332
c_mean_MFCC8	0.392	0.676	1.373	0.242	1.375	0.24	0.552	0.698
c_std_MFCC8	0.084	0.919	0.154	0.695	2.876	0.022	0.919	0.452
c_mean_MFCC6	0.018	0.982	0.936	0.334	2.035	0.087	1.011	0.4

c_mean_d2	1.782	0.169	0	1.0	0	1.0	6.742	<0.001
c_mean_d6	0.515	0.598	0	1.0	0	1.0	3.16	0.013
c_mean_MFCC3	1.393	0.249	1.886	0.17	7.87	<0.001	6.754	<0.001
c_mean_d1	1.664	0.19	0	1.0	0	1.0	7.083	<0.001

PD- Parkinson's Disease, HC- Healthy Control, c_mean- Mean of the MFCCs Coefficients, log- Energy of the Signal and the First and Second Derivatives of the MFCCs, MFCC- Mel Frequency Cepstral Coefficients, c_std- Standard Deviation of the MFCCs Coefficients, gqc- Glottis Quotient Close, fm- Frequency Modulation, std - Standard Deviation, tkeo- Teager Kaiser Energy Operator, p5- 5th percentile, p75- 75th Percentile, p95- 95th Percentile, shbd- Shimmer, hnr- Harmonic to Noise Ratio, d- Delta, d-d- Delta-Delta, DFA- Detrended Fluctuation Analysis, f0- Fundamental Frequency, ApEn- Pitch Period Entropy, rpde- Recurrence Period Density Entropy, T- Period.

Table S13. Results from an analysis of ANOVA for repeated measurements on the features from Tapping task. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	Medication		Diagnosis (PD, HC)		Repetition		Diagnosis × Repetition	
	F	*P*	F	*P*	F	*P*	F	*P*
numberTaps	4.903	0.008	539.151	<0.001	123.309	<0.001	4.444	0.001
max_TapInter	0.784	0.457	296.6	<0.001	4.609	0.001	1.233	0.294
range_TapInter	0.768	0.464	271.738	<0.001	3.207	0.012	1.308	0.264
ar2_TapInter	0.219	0.803	159.025	<0.001	6.706	<0.001	2.583	0.035
ar1_TapInter	0.194	0.824	268.543	<0.001	10.484	<0.001	3.829	0.004
sd_TapInter	1	0.369	212.428	<0.001	3.804	0.004	1.424	0.223
buttonNoneFreq	0.073	0.929	50.866	<0.001	4.93	<0.001	2.344	0.052
mad_TapInter	0.524	0.593	87.645	<0.001	0.721	0.578	4.172	0.002
median_DriftRight	0.128	0.88	8.987	0.003	9.747	<0.001	6.351	<0.001
mad_DriftRight	0.167	0.846	8.9	0.003	10.605	<0.001	6.353	<0.001
median_DriftLeft	0.382	0.683	4.491	0.034	3.107	0.015	2.769	0.026
min_TapInter	1.49	0.226	44.616	<0.001	11.437	<0.001	0.388	0.818
sd_DriftRight	1.279	0.279	4.338	0.037	6.689	<0.001	3.516	0.007
iqr_TapInter	1.145	0.319	53.591	<0.001	1.054	0.378	2.553	0.037
mad_DriftLeft	0.603	0.548	2.594	0.108	11.145	<0.001	6.06	<0.001
sd_DriftLeft	0.297	0.743	4.586	0.032	3.469	0.008	2.217	0.065
skew_TapInter	2.064	0.128	77.596	<0.001	1.117	0.346	2.171	0.07

skew_DriftLeft	2.606	0.075	33.216	<0.001	1.021	0.395	0.684	0.603
kur_DriftLeft	1.888	0.152	29.675	<0.001	1.196	0.31	0.414	0.799
kur_DriftRight	1.34	0.262	55.612	<0.001	2.173	0.069	1.553	0.184
kur_TapInter	1.672	0.189	45.738	<0.001	0.864	0.484	0.88	0.475
cv_TapInter	1.466	0.232	8.054	0.005	6.815	<0.001	1.702	0.146
corXY	0.135	0.874	5.935	0.015	21.815	<0.001	2.541	0.038
tkeo_TapInter	0.373	0.689	26.888	<0.001	0.424	0.792	0.388	0.817
cv_DriftLeft	2.424	0.089	0.005	0.944	0.573	0.682	1.793	0.127

PD- Parkinson's Disease, HC- Healthy Control, 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 Absolute Deviation, min- Minimum, max- Maximum, skew- Skewness, kur- Kurtosis, teko-Teager-Kaiser Energy Operator, cv- Coefficient, Sd- Standard Deviation, ar (1-2)- Coefficient of an Autoregressive Model at Lag (1-2).

Table S14. Results from an analysis of ANOVA for repeated measurements on the features from Gait task controlling for age and sex covariates. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	diagnosis (PD, HC)		Repetition		Diagnosis × Repetition	
	F	*P*	F	*P*	F	*P*
frec_peak_LB_a	6.944	0.008	1.952	0.163	0.993	0.319
FreezeInd_z	7.058	0.008	5.454	0.02	3.621	0.057
iqr_x	0.036	0.85	15.089	<0.001	0.041	0.84
MSI	3.067	0.08	0.45	0.502	0.017	0.897
numSteps	2.087	0.149	0.603	0.438	0.016	0.899
median_acc	2.8	0.094	0.024	0.877	0.408	0.523
PeakEnerg_LB_x	1.405	0.236	12.693	<0.001	0.008	0.927
min_z	11.446	0.001	0.621	0.431	5.301	0.021
ApEn_pos_z	5.645	0.018	0.435	0.51	2.665	0.103
Power_LB_x	0.832	0.362	7.791	0.005	0.001	0.974
Power_FB_z	1.26	0.262	22.277	<0.001	4.67	0.031
ApEn_pos_a	0.327	0.567	2.596	0.107	0.207	0.649
mean_a	0.091	0.762	14.784	<0.001	0.961	0.327
iqr_acc	0.228	0.633	16.093	<0.001	2.586	0.108

mean_acc	2.736	0.098	4.55	0.033	0.79	0.374
rms_acc	0.367	0.545	8.371	0.004	1.099	0.295
rms_y	0.47	0.493	5.649	0.018	1.738	0.188
PeakEnerg_LB_a	0.222	0.637	11.596	0.001	0.531	0.466
RatioPower_y	0.035	0.851	8.374	0.004	1.148	0.284
kur_pos_z	5.757	0.016	1.125	0.289	3.633	0.057
frec_peak_FB_vel_10	7.657	0.006	0.503	0.478	2.496	0.114
iqr_z	0.054	0.816	5.862	0.016	0.353	0.552
skew_z	1.17	0.28	3.897	0.048	0.053	0.818
frec_peak_FB_a	2.47	0.116	0.422	0.516	0.221	0.638
median_a	1.059	0.304	11.019	0.001	0.11	0.74
FreezeInd_x	0.696	0.404	0.046	0.831	0.227	0.633
Power_LB_z	6.087	0.014	7.609	0.006	0.134	0.715
max_acc	3.393	0.066	0.688	0.407	1.754	0.186
zcr_x	0.341	0.559	3.604	0.058	0	0.99
Power_LB_y	0.701	0.402	2.285	0.131	0.111	0.739
Power_FB_acc	0.008	0.927	11.015	0.001	0.7	0.403
kur_pos_a	0.242	0.623	6.086	0.014	0.566	0.452
iqr_y	0.005	0.946	1.366	0.243	0.276	0.599
FreezeInd_a	0.266	0.606	2.779	0.096	0.332	0.565
Power_FB_vel	0.001	0.971	0.545	0.461	0.001	0.975
kur_x	0.434	0.51	0.295	0.587	1.733	0.188
frec_peak_LB_z	2.132	0.144	7.517	0.006	1.52	0.218
COEFCEPS20_pos_y	0.592	0.442	2.244	0.134	0.697	0.404
COEFCEPS7_z	4.786	0.029	4.437	0.035	7.071	0.008
PeakEnerg_FB_x	1.881	0.17	8.074	0.005	0.076	0.783
skew_vel	1.274	0.259	3.714	0.054	0.369	0.543
cov_vel	3.082	0.079	0.861	0.354	0.162	0.687
COEFCEPS1_pos_x	2.37	0.124	2.696	0.101	0.043	0.836
iqr_pos_z	1.162	0.281	10.471	0.001	0.629	0.428
cov_a	1.337	0.248	0.003	0.958	1.068	0.302
FreezeInd_pos_a	0.655	0.418	4.943	0.026	2.571	0.109
COEFCEPS9_z	3.211	0.073	1.042	0.307	2.587	0.108

PeakEnerg_LB_z	4.942	0.026	5.678	0.017	0.342	0.559
COEFCEPS8_z	4.642	0.031	3.393	0.066	4.737	0.03
COEFCEPS6_z	3.364	0.067	3.745	0.053	4.44	0.035
frec_peak_LB_y	0.108	0.742	2.233	0.135	0.017	0.895
ApEn_vel	0.47	0.493	2.22	0.136	0.527	0.468
ApEn_pos_x	1.058	0.304	0.527	0.468	0.003	0.958
median_y	0.04	0.842	4.124	0.042	0.055	0.815
COEFCEPS10_z	1.739	0.187	0.256	0.613	1.947	0.163
PeakEnerg_LB_pos_z	1.927	0.165	6.041	0.014	0.454	0.5
COEFCEPS1_pos_a	2.811	0.094	0.217	0.641	1.267	0.26
zcr_pos_z	0.51	0.475	7.114	0.008	0.962	0.327
RatioPower_pos_z	0.811	0.368	5.956	0.015	0.14	0.708
RatioPower_pos_a	0.811	0.368	11.062	0.001	0.479	0.489
Power_FB_pos_z	0.647	0.421	6.161	0.013	0.116	0.733
ApEn_pos_y	0.363	0.547	2.754	0.097	0.099	0.753
kur_vel	0.223	0.636	0.957	0.328	0.621	0.431
cov_acc	0.078	0.781	0.001	0.976	0.092	0.762
min_pos_x	0.001	0.979	3.915	0.048	0.132	0.716
min_pos_y	1.527	0.217	10.666	0.001	1.791	0.181

PD- Parkinson's Disease, HC- Healthy Control, frec_peak- Frequency at the Peak of Energy, FreezeInd- Freeze Index, iqr-Interquartile Range, MSI- Mean Stride Interval, numSteps- Number of Steps, PeakEnerg- Peak of Energy, ApEn- Entropy, rms- Root Mean Square, RatioPower - Sum of the Power in the Freezing and Locomotor Band, skew- Skewness, min- Minimum Value, cov- Coefficient of Variation, zcr- Zero-Crossing Rate, kur-Kurtosis, COEFCEPS (1-20)- Mel Frequency Cepstral Coefficients, ar- Coefficient of a 1st Order Autoregressive Model, LB- Locomotor Band, FB- Freezing Band, vel- Velocity, acc- Acceleration Along Path, a- Accelerometer Average Signal, x- Accelerometer Mediolateral Signal, y- Accelerometer Vertical Signal, z- Accelerometer Anteroposterior Signal.

Table S15. Results from an analysis of ANOVA for repeated measurements on the features from Balance task controlling for age and sex covariates. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	Diagnosis (PD, HC)		Repetition		diagnosis × Repetition	
	F	*P*	F	*P*	F	*P*
Power_MF_trem_z	13.924	<0.001	1.713	0.191	0.442	0.506

PeakEnerg_VHF_trem_x	9.384	0.002	0.469	0.494	0.259	0.611
PeakEnerg_VHF_trem_z	6.816	0.009	0.168	0.682	0.106	0.745
RHL_trem_z	1.821	0.177	0.127	0.721	1.904	0.168
RHL_trem_a	9.039	0.003	0.254	0.614	0.24	0.624
Power_trem_y	13.351	<0.001	3.324	0.068	0.653	0.419
F95_post_y	1.028	0.311	4.288	0.038	0.011	0.916
Power_trem_z	9.023	0.003	4.007	0.045	0.363	0.547
median_trem_a	11.706	0.001	0.562	0.454	0.01	0.92
Power_LF_trem_z	2.945	0.086	7.861	0.005	0.267	0.606
CFREQ_post_z	7.946	0.005	0.048	0.826	2.48	0.115
F95_post_x	0.03	0.863	3.62	0.057	0.032	0.859
MFREQ_dist_x	3.866	0.049	0.051	0.821	0.041	0.839
iqr_post_y	12.53	<0.001	6.415	0.011	1.226	0.268
mean_trem_a	10.042	0.002	2.076	0.15	0.001	0.98
kur_trem_x	0.68	0.41	0.024	0.878	0.471	0.492
F95_post_a	0.473	0.492	0.561	0.454	0.031	0.86
ApEn_trem_x	1.082	0.298	1.172	0.279	0.091	0.763
zcr_post_y	4.644	0.031	2.524	0.112	0.001	0.971
median_post_a	12.753	<0.001	1.11	0.292	0.959	0.327
iqr_trem_x	1.52	0.218	2.875	0.09	0.163	0.686
FD_CC_dist_x_z	5.802	0.016	1.641	0.2	1.19	0.275
F50_post_y	0.082	0.775	11.891	0.001	2.087	0.149
Power_LF_trem_x	3.671	0.055	6.652	0.01	1.175	0.279
range_trem_y	10.445	0.001	4.427	0.035	1.457	0.228
MVELO_dist_x	2.023	0.155	16.408	<0.001	0.043	0.835
mean_post_y	0.169	0.681	21.508	<0.001	0.221	0.638
min_post_y	6.601	0.01	10.242	0.001	0.389	0.533
iqr_post_x	2.773	0.096	8.123	0.004	0.196	0.658
kur_post_x	0.667	0.414	0.241	0.624	0.056	0.813
rms_trem_a	7.56	0.006	1.993	0.158	0	0.986
ApEn_post_a	0.019	0.89	0.304	0.581	0.949	0.33

skew_post_a	1.915	0.166	0.243	0.622	0.082	0.775
AREA_CC_dist_x_z	6.846	0.009	0.303	0.582	0.691	0.406
FD_dist_x_z	5.323	0.021	2.363	0.124	2.494	0.114
cov_trem_a	1.552	0.213	0.274	0.6	0	0.994
TotalPower_post_x_z	0.701	0.403	0	0.995	0	0.988
MFREQ_dist_x_z	3.421	0.064	0.81	0.368	0.364	0.546
max_post_y	9.211	0.002	6.602	0.01	1.224	0.269
F50_post_x	1.008	0.315	2.508	0.113	1.285	0.257
cov_post_a	1.915	0.166	0.238	0.626	0	0.987
range_trem_x	0.846	0.358	4.672	0.031	0.925	0.336
ApEn_trem_y	0.042	0.838	0.031	0.861	1.655	0.198
FD_CE_dist_x_z	0.222	0.638	1.003	0.317	0.596	0.44
mean_post_z	2.909	0.088	0.82	0.365	3.05	0.081
F50_post_a	0.002	0.966	0.028	0.867	0.001	0.974
Power_LF_trem_a	1.369	0.242	0.167	0.683	0.003	0.958
max_post_z	0.452	0.502	13.973	<0.001	0.765	0.382
rms_post_a	7.417	0.007	2.16	0.142	0.535	0.464
F50_post_x_z	0.012	0.913	0.03	0.862	0.007	0.931
FREQD_post_x	0.414	0.52	2.813	0.094	2.632	0.105
TotalPower_post_z	0.251	0.616	5.167	0.023	0.102	0.75
FREQD_post_x_z	0.271	0.602	0.259	0.611	0.205	0.65
TotalPower_post_y	5.136	0.024	4.032	0.045	0.116	0.733
kur_post_a	2.183	0.14	0.604	0.437	0.07	0.791
max_trem_z	0.366	0.545	3.539	0.06	0.49	0.484
jerk_post_y	4.462	0.035	2.007	0.157	0.247	0.619
kur_trem_a	1.351	0.245	0.26	0.61	0.286	0.593
kur_trem_y	1.349	0.246	0.001	0.977	0.433	0.511

PD- Parkinson's Disease, HC- Healthy Control, PeakEnergy - Peak of energy, TotalPower- Energy between 15-3.5 Hz, rms- Root Mean Square, F50- Frequency Containing 50% of Total Power, F95- Frequency containing 95% of the total power, FRQD- Frequency of Dispersion of the Power Spectrum, MFREQ- Mean Frequency, iqr- Interquartile Range, kur- Kurtosis, zcr- Zero-Crossing Rate, ApEn- Entropy, skew- Skewness, jerk- Average jerk, MVELO- Mean velocity, FD- Fractal

Dimension, FD_CE- Fractal Dimension based on the 95% Confidence Ellipse Area, min- Minimum Value, CFREQ- Centroidal Frequency, RHL- Ratio Between Power in High Frequency and Low Frequency, dist- Distance, 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.

Table S16. Results from an analysis of ANOVA for repeated measurements on the features from Voice task controlling for age and sex covariates. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	Diagnosis (PD, HC)		Repetition		Diagnosis × Repetition	
	F	P	F	P	F	P
mean_gqc	2.682	0.102	20.802	<0.001	9.344	0.002
p5_gqc	3.736	0.053	0.638	0.425	0.01	0.921
c_mean_MFCC1	19.933	<0.001	17.793	<0.001	5.558	0.018
p95_gqc	0.263	0.608	10.682	0.001	5.969	0.015
std_tkeo	0.005	0.942	0.215	0.643	2.366	0.124
p95_tkeo	0.075	0.785	0.448	0.503	0.996	0.318
c_std_d11	2.403	0.121	8.726	0.003	5.208	0.023
hnr_std	3.659	0.056	8.595	0.003	3.808	0.051
c_std_d12	6.39	0.012	19.707	<0.001	14.38	<0.001
p75_tkeo	0.544	0.461	0.035	0.852	3.077	0.079
c_std_d13	4.411	0.036	9.512	0.002	8.334	0.004
c_std_d8	2.207	0.137	6.708	0.01	6.044	0.014
c_std_d9	3.096	0.079	12.871	<0.001	5.54	0.019
c_mean_MFCC10	3.742	0.053	0.892	0.345	0.147	0.702
c_std_d10	2.394	0.122	11.998	0.001	7.408	0.007
c_std_d7	3.362	0.067	12.585	<0.001	8.865	0.003
c_std_dd11	1.264	0.261	8.035	0.005	7.171	0.007
c_std_d5	4.501	0.034	6.582	0.01	4.272	0.039
c_std_d14	3.095	0.079	9.011	0.003	6.748	0.009
c_std_d6	4.115	0.043	8.333	0.004	4.673	0.031
c_mean_MFCC12	0.039	0.843	0.958	0.328	0.021	0.884
c_std_dd5	2.417	0.12	6.101	0.014	4.911	0.027

c_std_d3	10.323	0.001	3.912	0.048	5.943	0.015
c_std_d4	2.603	0.107	1.899	0.168	1.285	0.257
c_std_dd10	0.896	0.344	9.493	0.002	8.42	0.004
c_std_dd9	1.996	0.158	13.165	<0.001	8.368	0.004
c_std_dd8	1.061	0.303	7.266	0.007	7.507	0.006
c_std_dd12	2.707	0.1	15.194	<0.001	12.525	<0.001
c_std_MFCC1	4.081	0.043	8.317	0.004	3.379	0.066
c_std_MFCC5	0.458	0.499	0.054	0.816	6.062	0.014
c_std_dd7	1.648	0.199	12.355	<0.001	10.179	0.001
c_std_dd13	2.037	0.154	9.669	0.002	9.633	0.002
c_std_dd6	2.063	0.151	8.666	0.003	6.784	0.009
DFA	5.291	0.021	2.569	0.109	0.686	0.408
c_mean_0th	0.004	0.952	63.748	<0.001	0.271	0.602
c_mean_d13	2.171	0.141	0.185	0.667	1.415	0.234
fm	0.006	0.94	2.907	0.088	1.257	0.262
c_std_dd4	0.986	0.321	2.298	0.13	1.526	0.217
c_mean_MFCC7	0.178	0.673	1.267	0.26	4.599	0.032
c_mean_d4	0.562	0.453	26.502	<0.001	8.242	0.004
shdb	0.005	0.944	0.087	0.769	0.595	0.441
c_std_dd3	8.805	0.003	8.041	0.005	10.768	0.001
ApEn_f0	0.025	0.875	5.173	0.023	0.089	0.765
c_std_dd14	1.961	0.162	10.961	0.001	11.192	0.001
c_mean_d3	4.019	0.045	7.308	0.007	0.455	0.5
c_std_MFCC11	0.121	0.728	0.613	0.434	3.232	0.072
c_std_d1	2.081	0.149	0.033	0.856	3.754	0.053
c_std_MFCC6	0.446	0.504	0.55	0.458	3.634	0.057
c_mean_MFCC5	1.023	0.312	5.317	0.021	0.415	0.519
c_mean_MFCC9	0.676	0.411	3.936	0.047	0.266	0.606
c_std_MFCC9	0.258	0.611	2.172	0.141	0.693	0.405
rpde	1.348	0.246	55.031	<0.001	3.912	0.048
c_std_MFCC7	0.064	0.801	1.815	0.178	2.478	0.116
c_mean_MFCC8	4.007	0.045	2.391	0.122	0.23	0.632
c_std_MFCC8	0.125	0.724	0.585	0.444	1.576	0.209

c_mean_MFCC6	4.507	0.034	7.557	0.006	3.609	0.058
c_mean_d2	0.486	0.486	6.366	0.012	7.916	0.005
c_mean_d6	0.428	0.513	4.818	0.028	4.267	0.039
c_mean_MFCC3	5.996	0.014	45.224	<0.001	16.228	<0.001
c_mean_d1	0.372	0.542	8.089	0.004	8.215	0.004

PD- Parkinson's Disease, HC- Healthy Control, c_mean- Mean of the MFCCs Coefficients, log- Energy of the Signal and the First and Second Derivatives of the MFCCs, MFCC- Mel Frequency Cepstral Coefficients, c_std- Standard Deviation of the MFCCs Coefficients, gqc- Glottis Quotient Close, fm- Frequency Modulation, std - Standard Deviation, tkeo- Teager Kaiser Energy Operator, p5- 5th percentile, p75- 75th Percentile, p95- 95th Percentile, shbd- Shimmer, hnr- Harmonic to Noise Ratio, d- Delta, d-d- Delta-Delta, DFA- Detrended Fluctuation Analysis, f0- Fundamental Frequency, ApEn- Pitch Period Entropy, rpde- Recurrence Period Density Entropy, T- Period.

Table S17. Results from an analysis of ANOVA for repeated measurements on the features from Tapping task controlling for age and sex covariates. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (*P*<.05).

Feature Name	Diagnosis (PD, HC)		Repetition		Diagnosis × Repetition	
	F	*P*	F	*P*	F	*P*
numberTaps	173.949	<0.001	270.014	<0.001	4.113	0.043
max_TapInter	87.799	<0.001	3.689	0.055	0.731	0.393
range_TapInter	80.469	<0.001	1.652	0.199	0.92	0.338
ar2_TapInter	64.885	<0.001	22.39	<0.001	4.956	0.026
ar1_TapInter	70.435	<0.001	18.438	<0.001	0.414	0.52
sd_TapInter	69.546	<0.001	2.817	0.093	0.345	0.557
buttonNoneFreq	19.865	<0.001	12.57	<0.001	1.617	0.204
mad_TapInter	25.648	<0.001	3.08	0.079	1.046	0.306
median_DriftRight	16.128	<0.001	54.313	<0.001	20.345	<0.001
mad_DriftRight	15.047	<0.001	52.104	<0.001	16.997	<0.001
median_DriftLeft	9.186	0.002	14.847	<0.001	9.512	0.002
min_TapInter	10.072	0.002	29.031	<0.001	0.952	0.329
sd_DriftRight	7.7	0.006	36.815	<0.001	12.629	<0.001
iqr_TapInter	13.74	<0.001	4.417	0.036	1.163	0.281
mad_DriftLeft	9.532	0.002	53.309	<0.001	21.249	<0.001
sd_DriftLeft	3.802	0.051	13.604	<0.001	4.779	0.029

skew_TapInter	13.569	<0.001	0.109	0.742	0.322	0.57
skew_DriftLeft	10.641	0.001	0.329	0.566	0.543	0.461
kur_DriftLeft	7.054	0.008	0.056	0.813	0.159	0.69
kur_DriftRight	5.379	0.02	10.3	0.001	3.462	0.063
kur_TapInter	6.747	0.009	0.796	0.372	0.001	0.978
cv_TapInter	1.777	0.183	7.905	0.005	1.254	0.263
corXY	12.801	<0.001	39.047	<0.001	3.865	0.049
tkeo_TapInter	2.802	0.094	0.042	0.837	0.571	0.45
cv_DriftLeft	5.924	0.015	4.297	0.038	3.211	0.073

PD- Parkinson's Disease, HC- Healthy Control, 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 Absolute Deviation, min- Minimum, max- Maximum, skew- Skewness, kur- Kurtosis, teko-Teager-Kaiser Energy Operator, cv- Coefficient, Sd- Standard Deviation, ar (1-2)- Coefficient of an Autoregressive Model at Lag (1-2).

Table S18. Results from an analysis of rm-ANOVA for elapse-time between repetition on the features from Gait task with controlling for age and sex covariates. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (P<.05).

Feature Name	diagnosis (PD, HC)		Elapsed-time		Diagnosis × Elapsed-time	
	F	*P*	F	*P*	F	*P*
frec_peak_LB_a	6.944	0.008	0.587	0.444	0.514	0.474
FreezeInd_z	7.058	0.008	2.381	0.123	4.074	0.044
iqr_x	0.036	0.85	0.468	0.494	2.821	0.093
MSI	3.067	0.08	0.027	0.87	0.312	0.577
numSteps	2.087	0.149	0.052	0.819	0.429	0.513
median_acc	2.8	0.094	1.017	0.313	2.433	0.119
PeakEnerg_LB_x	1.405	0.236	0.238	0.626	1.68	0.195
min_z	11.446	0.001	0.038	0.846	0.03	0.862
ApEn_pos_z	5.645	0.018	0.024	0.876	1.011	0.315
Power_LB_x	0.832	0.362	0.908	0.341	3.932	0.047
Power_FB_z	1.26	0.262	0.033	0.855	0.044	0.833
ApEn_pos_a	0.327	0.567	2.002	0.157	2.41	0.121
mean_a	0.091	0.762	0.869	0.351	1.93	0.165

iqr_acc	0.228	0.633	0.102	0.749	0.161	0.688
mean_acc	2.736	0.098	0.029	0.864	0.132	0.716
rms_acc	0.367	0.545	0.079	0.778	0.284	0.594
rms_y	0.47	0.493	0.201	0.654	1.118	0.29
PeakEnerg_LB_a	0.222	0.637	0.101	0.751	0.197	0.657
RatioPower_y	0.035	0.851	0.157	0.692	1.339	0.247
kur_pos_z	5.757	0.016	0.08	0.777	0.514	0.474
frec_peak_FB_vel	7.657	0.006	1.89	0.169	2.779	0.096
iqr_z	0.054	0.816	2.288	0.13	2.123	0.145
skew_z	1.17	0.28	0.327	0.567	0.001	0.981
frec_peak_FB_a	2.47	0.116	2.913	0.088	0.279	0.598
median_a	1.059	0.304	1.682	0.195	2.464	0.117
FreezeInd_x	0.696	0.404	1.935	0.164	4.066	0.044
Power_LB_z	6.087	0.014	0.255	0.614	1.424	0.233
max_acc	3.393	0.066	0.162	0.687	0.821	0.365
zcr_x	0.341	0.559	0.023	0.88	3.326	0.068
Power_LB_y	0.701	0.402	0.585	0.444	2.772	0.096
Power_FB_acc	0.008	0.927	0.637	0.425	0.143	0.705
kur_pos_a	0.242	0.623	1.089	0.297	1.446	0.229
iqr_y	0.005	0.946	1.297	0.255	2.27	0.132
FreezeInd_a	0.266	0.606	0	0.989	0.507	0.477
Power_FB_vel	0.001	0.971	1.025	0.311	1.475	0.225
kur_x	0.434	0.51	0.112	0.738	1.005	0.316
frec_peak_LB_z	2.132	0.144	0.542	0.462	0.213	0.644
COEFCEPS20_pos_y	0.592	0.442	1.671	0.196	2.712	0.1
COEFCEPS7_z	4.786	0.029	0.004	0.95	0.048	0.827
PeakEnerg_FB_x	1.881	0.17	0.107	0.744	0.044	0.833
skew_vel	1.274	0.259	0.352	0.553	1.082	0.298
cov_vel	3.082	0.079	0.124	0.725	0.335	0.563
COEFCEPS1_pos_x	2.37	0.124	0.133	0.715	0.998	0.318
iqr_pos_z	1.162	0.281	0.456	0.5	0.007	0.933
cov_a	1.337	0.248	1.743	0.187	1.226	0.268
FreezeInd_pos_a	0.655	0.418	0.058	0.81	0.204	0.651

COEFCEPS9_z	3.211	0.073	0.194	0.659	0.033	0.856
PeakEnerg_LB_z	4.942	0.026	0.969	0.325	1.917	0.166
COEFCEPS8_z	4.642	0.031	0.367	0.545	0.301	0.584
COEFCEPS6_z	3.364	0.067	0	0.988	0.032	0.859
frec_peak_LB_y	0.108	0.742	0	0.994	0.113	0.737
ApEn_vel	0.47	0.493	0.531	0.466	1.529	0.216
ApEn_pos_x	1.058	0.304	0.064	0.8	4.145	0.042
median_y	0.04	0.842	0.232	0.63	0.657	0.418
COEFCEPS10_z	1.739	0.187	4.545	0.033	3.519	0.061
PeakEnerg_LB_pos_z	1.927	0.165	0.886	0.347	0.123	0.726
COEFCEPS1_pos_a	2.811	0.094	1.017	0.313	2.23	0.135
zcr_pos_z	0.51	0.475	0.2	0.654	0.019	0.889
RatioPower_pos_z	0.811	0.368	0.792	0.374	0.076	0.782
RatioPower_pos_a	0.811	0.368	1.351	0.245	0.384	0.535
Power_FB_pos_z	0.647	0.421	0.812	0.367	0.083	0.773
ApEn_pos_y	0.363	0.547	0.664	0.415	0.621	0.431
kur_vel	0.223	0.636	0.747	0.387	1.305	0.253
cov_acc	0.078	0.781	0.002	0.965	0.143	0.706
min_pos_x	0.001	0.979	0.025	0.876	0.012	0.913
min_pos_y	1.527	0.217	0.197	0.657	0.137	0.711

PD- Parkinson’s Disease, HC- Healthy Control, frec_peak- Frequency at the Peak of Energy, FreezeInd- Freeze Index, iqr-Interquartile Range, MSI- Mean Stride Interval, numSteps- Number of Steps, PeakEnerg- Peak of Energy, ApEn- Entropy, rms- Root Mean Square, RatioPower - Sum of the Power in the Freezing and Locomotor Band, skew- Skewness, min- Minimum Value, cov- Coefficient of Variation, zcr- Zero-Crossing Rate, kur-Kurtosis, COEFCEPS (1-20)- Mel Frequency Cepstral Coefficients, ar- Coefficient of a 1st Order Autoregressive Model, LB- Locomotor Band, FB- Freezing Band, vel- Velocity, acc- Acceleration Along Path, a- Accelerometer Average Signal, x- Accelerometer Mediolateral Signal, y- Accelerometer Vertical Signal, z- Accelerometer Anteroposterior Signal.

Table S29. Results from an analysis of rm-ANOVA for elapsed-time between repetition on the features from Balance task with controlling for age and sex covariates. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (P<.05).

Feature Name	Diagnosis (PD, HC)		Elapsed-time		diagnosis × Elapsed-time	
	F	*P*	F	*P*	F	*P*

Power_MF_trem_z	13.924	<0.001	0.854	0.355	1.034	0.309
PeakEnerg_VHF_trem_x	9.384	0.002	0.005	0.942	0.752	0.386
PeakEnerg_VHF_trem_z	6.816	0.009	0.012	0.912	0.303	0.582
RHL_trem_z	1.821	0.177	0.008	0.93	0.019	0.891
RHL_trem_a	9.039	0.003	0.187	0.666	0.284	0.594
Power_trem_y	13.351	<0.001	0.795	0.373	0.39	0.533
F95_post_y	1.028	0.311	2.783	0.095	2.708	0.1
Power_trem_z	9.023	0.003	2.267	0.132	3.226	0.073
median_trem_a	11.706	0.001	0.431	0.512	0.009	0.925
Power_LF_trem_z	2.945	0.086	3.484	0.062	1.019	0.313
CFREQ_post_z	7.946	0.005	0.522	0.47	0.136	0.712
F95_post_x	0.03	0.863	0.8	0.371	2.23	0.135
MFREQ_dist_x	3.866	0.049	0.472	0.492	0.325	0.568
iqr_post_y	12.53	<0.001	0.177	0.674	1.407	0.236
mean_trem_a	10.042	0.002	1.606	0.205	0.125	0.724
kur_trem_x	0.68	0.41	0.128	0.72	0.022	0.883
F95_post_a	0.473	0.492	0.734	0.392	0.035	0.852
ApEn_trem_x	1.082	0.298	0.156	0.693	0.128	0.72
zcr_post_y	4.644	0.031	0.231	0.631	0.848	0.357
median_post_a	12.753	<0.001	3.774	0.052	2.274	0.132
iqr_trem_x	1.52	0.218	0.712	0.399	0.114	0.736
FD_CC_dist_x_z	5.802	0.016	0.006	0.941	0.102	0.75
F50_post_y	0.082	0.775	13.787	<0.001	9.119	0.003
Power_LF_trem_x	3.671	0.055	0	0.998	1.853	0.174
range_trem_y	10.445	0.001	0.562	0.453	0.284	0.594
MVELO_dist_x	2.023	0.155	0.559	0.455	0.016	0.899
mean_post_y	0.169	0.681	2.854	0.091	2.325	0.127
min_post_y	6.601	0.01	0.24	0.624	1.263	0.261
iqr_post_x	2.773	0.096	1.81	0.179	0.601	0.438
kur_post_x	0.667	0.414	0.173	0.678	0.002	0.967
rms_trem_a	7.56	0.006	2.954	0.086	0.245	0.621

ApEn_post_a	0.019	0.89	0.207	0.65	0.072	0.788
skew_post_a	1.915	0.166	0.184	0.668	0.001	0.977
AREA_CC_dist_x_z	6.846	0.009	4.776	0.029	4.564	0.033
FD_dist_x_z	5.323	0.021	0.139	0.709	0.096	0.756
cov_trem_a	1.552	0.213	0.577	0.447	0.318	0.573
TotalPower_post_x_z	0.701	0.403	5.833	0.016	1.417	0.234
MFREQ_dist_x_z	3.421	0.064	0.023	0.88	0	1
max_post_y	9.211	0.002	0.495	0.482	1.869	0.172
F50_post_x	1.008	0.315	0.488	0.485	0.09	0.764
cov_post_a	1.915	0.166	1.204	0.273	0.451	0.502
range_trem_x	0.846	0.358	0.17	0.68	0.143	0.705
ApEn_trem_y	0.042	0.838	4.86	0.028	5.292	0.021
FD_CE_dist_x_z	0.222	0.638	0.824	0.364	0.018	0.892
mean_post_z	2.909	0.088	3.298	0.069	4.407	0.036
F50_post_a	0.002	0.966	2.051	0.152	0.404	0.525
Power_LF_trem_a	1.369	0.242	3.606	0.058	0.424	0.515
max_post_z	0.452	0.502	8.326	0.004	6.825	0.009
rms_post_a	7.417	0.007	6.689	0.01	3.242	0.072
F50_post_x_z	0.012	0.913	1.016	0.313	0.157	0.692
FREQD_post_x	0.414	0.52	1.239	0.266	0.249	0.618
TotalPower_post_z	0.251	0.616	7.714	0.006	3.893	0.049
FREQD_post_x_z	0.271	0.602	0.158	0.691	0.572	0.45
TotalPower_post_y	5.136	0.024	0.121	0.728	1.147	0.284
kur_post_a	2.183	0.14	0.023	0.878	0.01	0.92
max_trem_z	0.366	0.545	2.946	0.086	4.203	0.04
jerk_post_y	4.462	0.035	0.089	0.766	0.12	0.729
kur_trem_a	1.351	0.245	0.042	0.838	0.01	0.919
kur_trem_y	1.349	0.246	0.194	0.66	0.479	0.489

PD- Parkinson’s Disease, HC- Healthy Control, PeakEnergy - Peak of energy, TotalPower- Energy between 15-3.5 Hz, rms- Root Mean Square, F50- Frequency Containing 50% of Total Power, F95- Frequency containing 95% of the total power, FRQD- Frequency of Dispersion of the Power

Spectrum, MFREQ- Mean Frequency, iqr- Interquartile Range, kur- Kurtosis, zcr- Zero-Crossing Rate, ApEn- Entropy, skew- Skewness, jerk- Average jerk, MVELO- Mean velocity, FD- Fractal Dimension, FD_CE- Fractal Dimension based on the 95% Confidence Ellipse Area, min- Minimum Value, CFREQ- Centroidal Frequency, RHL- Ratio Between Power in High Frequency and Low Frequency, dist- Distance, 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.

Table S20. Results from an analysis of rm-ANOVA for elapsed-time between repetition on the features from Voice task with controlling for age and sex covariates. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (P<.05).

Feature Name	Diagnosis (PD, HC)		Elapsed-time		Diagnosis × Elapsed-time	
	F	P	F	P	F	P
mean_gqc	2.682	0.102	0.864	0.353	0.33	0.566
p5_gqc	3.736	0.053	0.43	0.512	0.13	0.719
c_mean_MFCC1	19.933	<0.001	0.222	0.637	3.691	0.055
p95_gqc	0.263	0.608	0.004	0.95	0.01	0.921
std_tkeo	0.005	0.942	0.707	0.401	0.011	0.915
p95_tkeo	0.075	0.785	0.264	0.607	0.073	0.786
c_std_d11	2.403	0.121	2.202	0.138	0.06	0.806
hnr_std	3.659	0.056	0.111	0.739	0.046	0.831
c_std_d12	6.39	0.012	5.903	0.015	2.081	0.149
p75_tkeo	0.544	0.461	0.821	0.365	0.417	0.519
c_std_d13	4.411	0.036	4.223	0.04	1.025	0.311
c_std_d8	2.207	0.137	4.926	0.027	1.566	0.211
c_std_d9	3.096	0.079	3.292	0.07	0.379	0.538
c_mean_MFCC10	3.742	0.053	0.786	0.375	0.002	0.961
c_std_d10	2.394	0.122	5.104	0.024	1.247	0.264
c_std_d7	3.362	0.067	4.4	0.036	1.44	0.23
c_std_dd11	1.264	0.261	1.162	0.281	0.249	0.618
c_std_d5	4.501	0.034	3.478	0.062	1.019	0.313
c_std_d14	3.095	0.079	1.319	0.251	0.483	0.487
c_std_d6	4.115	0.043	2.967	0.085	0.698	0.404

c_mean_MFCC12	0.039	0.843	1.216	0.27	0.082	0.775
c_std_dd5	2.417	0.12	2.163	0.141	0.482	0.488
c_std_d3	10.323	0.001	0.543	0.461	0.072	0.788
c_std_d4	2.603	0.107	0.538	0.463	0	0.989
c_std_dd10	0.896	0.344	2.078	0.149	0.234	0.628
c_std_dd9	1.996	0.158	2.366	0.124	0.034	0.854
c_std_dd8	1.061	0.303	3.2	0.074	0.784	0.376
c_std_dd12	2.707	0.1	3.411	0.065	0.646	0.422
c_std_MFCC1	4.081	0.043	0.091	0.762	0.511	0.475
c_std_MFCC5	0.458	0.499	0.019	0.889	0	0.996
c_std_dd7	1.648	0.199	2.823	0.093	0.428	0.513
c_std_dd13	2.037	0.154	3.453	0.063	0.653	0.419
c_std_dd6	2.063	0.151	1.882	0.17	0.717	0.397
DFA	5.291	0.021	0.388	0.534	0.225	0.635
c_mean_0th	0.004	0.952	6.935	0.008	1.78	0.182
c_mean_d13	2.171	0.141	0.01	0.922	0.211	0.646
fm	0.006	0.94	0.34	0.56	0.264	0.607
c_std_dd4	0.986	0.321	0.293	0.588	0.032	0.858
c_mean_MFCC7	0.178	0.673	0.661	0.416	2.161	0.142
c_mean_d4	0.562	0.453	1.98	0.159	0.006	0.939
shdb	0.005	0.944	0.248	0.618	0.001	0.972
c_std_dd3	8.805	0.003	0.82	0.365	0.273	0.602
ApEn_f0	0.025	0.875	2.595	0.107	0.225	0.635
c_std_dd14	1.961	0.162	1.393	0.238	0.461	0.497
c_mean_d3	4.019	0.045	0.733	0.392	11.176	0.001
c_std_MFCC11	0.121	0.728	0.485	0.486	2.227	0.136
c_std_d1	2.081	0.149	0.013	0.908	0.312	0.577
c_std_MFCC6	0.446	0.504	0.521	0.47	2.602	0.107
c_mean_MFCC5	1.023	0.312	8.88	0.003	3.676	0.055
c_mean_MFCC9	0.676	0.411	7.4	0.007	1.474	0.225
c_std_MFCC9	0.258	0.611	0.243	0.622	1.114	0.291
rpde	1.348	0.246	7.508	0.006	1.289	0.256
c_std_MFCC7	0.064	0.801	0.03	0.863	0.118	0.731

c_mean_MFCC8	4.007	0.045	6.575	0.01	0.554	0.457
c_std_MFCC8	0.125	0.724	0.98	0.322	0.057	0.812
c_mean_MFCC6	4.507	0.034	5.187	0.023	0.884	0.347
c_mean_d2	0.486	0.486	0.971	0.324	1.459	0.227
c_mean_d6	0.428	0.513	1.936	0.164	3.473	0.062
c_mean_MFCC3	5.996	0.014	12.14	<0.001	2.928	0.087
c_mean_d1	0.372	0.542	1.09	0.296	1.326	0.249

PD- Parkinson's Disease, HC- Healthy Control, c_mean- Mean of the MFCCs Coefficients, log- Energy of the Signal and the First and Second Derivatives of the MFCCs, MFCC- Mel Frequency Cepstral Coefficients, c_std- Standard Deviation of the MFCCs Coefficients, gqc- Glottis Quotient Close, fm- Frequency Modulation, std - Standard Deviation, tkeo- Teager Kaiser Energy Operator, p5- 5th percentile, p75- 75th Percentile, p95- 95th Percentile, shbd- Shimmer, hnr- Harmonic to Noise Ratio, d- Delta, d-d- Delta-Delta, DFA- Detrended Fluctuation Analysis, f0- Fundamental Frequency, ApEn- Pitch Period Entropy, rpde- Recurrence Period Density Entropy, T- Period.

Table S21. Results from an analysis of rm-ANOVA for elapsed-time between repetition on the features from Tapping task with controlling for age and sex covariates. Features are selected based on Mann-Whitney U test that significantly differ between PD and HC at the first administration (baseline) (P<.05).

Feature Name	Diagnosis (PD, HC)		Elapsed-time		Diagnosis × Elapsed-time	
	F	*P*	F	*P*	F	*P*
numberTaps	173.949	<0.001	21.87	<0.001	0.039	0.844
max_TapInter	87.799	<0.001	0.589	0.443	0.124	0.725
range_TapInter	80.469	<0.001	0.534	0.465	0.094	0.76
ar2_TapInter	64.885	<0.001	3.462	0.063	0.832	0.362
ar1_TapInter	70.435	<0.001	1.603	0.206	0.026	0.873
sd_TapInter	69.546	<0.001	0.793	0.373	0.555	0.456
buttonNoneFreq	19.865	<0.001	2.614	0.106	0.994	0.319
mad_TapInter	25.648	<0.001	4.582	0.032	6.289	0.012
median_DriftRight	16.128	<0.001	4.486	0.034	1.344	0.246
mad_DriftRight	15.047	<0.001	3.759	0.053	1.349	0.246
median_DriftLeft	9.186	0.002	1.933	0.164	0.62	0.431
min_TapInter	10.072	0.002	0.056	0.813	0.477	0.49
sd_DriftRight	7.7	0.006	4.143	0.042	2.29	0.13
iqr_TapInter	13.74	<0.001	4.865	0.027	5.282	0.022

mad_DriftLeft	9.532	0.002	4.29	0.038	1.26	0.262
sd_DriftLeft	3.802	0.051	2.639	0.104	0.144	0.705
skew_TapInter	13.569	<0.001	0.147	0.702	0.164	0.686
skew_DriftLeft	10.641	0.001	0.008	0.93	2.264	0.132
kur_DriftLeft	7.054	0.008	0.352	0.553	1.516	0.218
kur_DriftRight	5.379	0.02	0.482	0.488	0.025	0.875
kur_TapInter	6.747	0.009	0.623	0.43	0.039	0.844
cv_TapInter	1.777	0.183	1.436	0.231	4.9	0.027
corXY	12.801	<0.001	3.85	0.05	0.787	0.375
tkeo_TapInter	2.802	0.094	0.002	0.962	3.785	0.052
cv_DriftLeft	5.924	0.015	0.025	0.875	1.907	0.167

PD- Parkinson's Disease, HC- Healthy Control, 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 Absolute Deviation, min- Minimum, max- Maximum, skew- Skewness, kur- Kurtosis, teko- Teager-Kaiser Energy Operator, cv- Coefficient, Sd- Standard Deviation, ar (1-2)- Coefficient of an Autoregressive Model at Lag (1-2).

Reference

1. Mirelman A, Heman T, Yasinovsky K, Thaler A, Gurevich T, Marder K, et al. Fall risk and gait in Parkinson's disease: the role of the LRRK2 G2019S mutation. Mov Disord 2013 Oct 7;28(12):1683–1690. PMID: 24123150
2. Sejdić E, Lowry KA, Bellanca J, Redfern MS, Brach JS. A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains. IEEE Trans Neural Syst Rehabil Eng 2014 May;22(3):603–612. PMID: 23751971
3. Zhan A, Little MA, Harris DA, Abiola SO, Dorsey ER, Saria S, et al. High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection. arXiv 2016 Jan 5;
4. Arora S, Venkataraman V, Zhan A, Donohue S, Biglan KM, Dorsey ER, et al. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. Parkinsonism Relat Disord 2015 Jun;21(6):650–653. PMID: 25819808
5. Weiss A, Sharifi S, Plotnik M, van Vugt JPP, Giladi N, Hausdorff JM. Toward automated, at-home assessment of mobility among patients with Parkinson disease, using a body-worn accelerometer. Neurorehabil Neural Repair 2011 Dec;25(9):810–818. PMID: 21989633
6. San-Segundo R, Torres-Sánchez R, Hodgins J, De la Torre F. Increasing robustness in the detection of freezing of gait in parkinson's disease. Electronics 2019 Jan 22;8(2):119.
7. Bächlin M, Plotnik M, Roggen D, Maidan I, Hausdorff JM, Giladi N, et al. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. IEEE Trans Inf Technol Biomed 2010 Mar;14(2):436–446. PMID: 19906597
8. 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 2011 May;15(3):481–490. PMID: 21349795

9. 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 2012;15(9):899–910. PMID: 21547782
10. Prieto TE, Myklebust JB, Hoffmann RG, Lovett EG, Myklebust BM. Measures of postural steadiness: differences between healthy young and elderly adults. IEEE Trans Biomed Eng 1996 Sep;43(9):956–966. PMID: 9214811
11. Tsanas A, Little MA, McSharry PE, Ramig LO. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. J R Soc Interface 2011 Jun 6;8(59):842–855. PMID: 21084338
12. Prince J, Andreotti F, De Vos M. Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-Wise Missing Data. IEEE Trans Biomed Eng 2019;66(5):1402–1411. PMID: 30403615
13. GitHub - Sage-Bionetworks/mPower-sdata: scripts used for preparation of Nature Scientific Data submission and release of mPower data [Internet]. [cited 2020 Aug 5]. Available from: https://github.com/Sage-Bionetworks/mPower-sdata/
14. Synapse | Sage Bionetworks [Internet]. [cited 2020 Aug 5]. Available from: https://www.synapse.org/#!Synapse:syn4993293/files/

Paper4: Smartphone-Based Digital Biomarkers for Parkinson's Disease in a Remotely Administered Setting

Maria Goni[1,2], Simon B Eickhoff[1,2], Mehran Sahandi Far[1,2], Kaustubh R Patil[1,2], Juergen Dukart[1,2]

[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

Corresponding Author:

Juergen Dukart, PhD
Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7) Research Centre Jülich
Wilhelm-Johnen-Strasse
Jülich, 52425
Germany
Phone: 49 1632874330
Fax: 49 2461611880
Email: juergen.dukart@gmail.com

Published in: *IEEE Access*, vol. 10, pp. 28361-28384, 2022, https://doi.org/10.1109/ACCESS.2022.3156659

Own contributions

Contributing and giving input on analyses, Own contribution 10%

Received January 27, 2022, accepted February 25, 2022, date of publication March 3, 2022, date of current version March 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3156659

Smartphone-Based Digital Biomarkers for Parkinson's Disease in a Remotely-Administered Setting

MARÍA GOÑI, SIMON B. EICKHOFF, MEHRAN SAHANDI FAR, KAUSTUBH R. PATIL, (Member, IEEE), AND JUERGEN DUKART

Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Jülich, 52425 Jülich, Germany
Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

Corresponding author: Juergen Dukart (juergen.dukart@gmail.com)

ABSTRACT Smartphone-based digital biomarker (DB) assessments provide objective measures of daily-life tasks and thus hold the promise to improve diagnosis and monitoring of Parkinson's disease (PD). To date, little is known about which tasks perform best for these purposes and how different confounds including comorbidities, age and sex affect their accuracy. Here we systematically assess the ability of common self-administered smartphone-based tasks to differentiate PD patients and healthy controls (HC) with and without accounting for the above confounds. Using a large cohort of PD patients and healthy volunteers acquired in the mPower study, we extracted about 700 features commonly reported in previous PD studies for gait, balance, voice and tapping tasks. We perform a series of experiments systematically assessing the effects of age, sex and comorbidities on the accuracy of the above tasks for differentiation of PD patients and HC using several machine learning algorithms. When accounting for age, sex and comorbidities, the highest balanced accuracy on hold-out data (73%) was achieved using random forest when combining all tasks followed by tapping using relevance vector machine (67%). Only moderate accuracies were achieved for other tasks (60% for balance, 56% for gait and 53% for voice data). Not accounting for the confounders consistently yielded higher accuracies of up to 77% when combining all tasks. Our results demonstrate the importance of controlling DB data for age and comorbidities.

INDEX TERMS Digital biomarkers, machine learning, Parkinson's disease, smartphones, wearable devices.

I. INTRODUCTION

Diagnosis of Parkinson's disease (PD) still often relies on in-clinic visits and evaluation based on clinical judgement as well as patient and caregiver reported information. This lack of objective measures and the need for in-clinic visits result in the often late and initially inaccurate diagnosis [1]. Recent studies have identified digital assessments as such promising objective biomarkers for PD symptoms including bradykinesia [2], [3], freezing of gait [4], [5], impaired dexterity [6], balance and speech difficulties [7]–[9]. Most of these results were obtained with a moderate number of participants and in a standardized and controlled clinical setting, reducing generalizability and limiting an interpretation with respect to applicability of these measures to an at-home self-administered setting [10]–[12].

As most relevant sensors deployed in these in-clinic studies are also embedded in modern smartphones, this opens the possibility to collect such objective, reliable and quantitative information as digital biomarkers (DB) in an at-home setting and therewith to facilitate diagnosis, health monitoring or treatment management using low-cost, simple and portable technology [13]. Indeed, recent studies applying machine learning algorithms to these high-dimensional data suggested a good diagnostic sensitivity of the respective digital assessments for detection of Parkinson's disease [14]–[17]. However, such at-home assessments create a range of new challenges including selection bias, confounding and sources of noise that need to be understood and dealt with to ensure good reliability of respective outcomes to a level that is sufficient for at home data collection [18]. For example, age, sex and comorbidities are known confounding factors that impact many measures of disease symptoms across neurodegenerative diseases including PD [19]–[23]. Yet, several

The associate editor coordinating the review of this manuscript and approving it for publication was Masood Ur-Rehman.

TABLE 1. Demographics for PD and HC subjects for each experiment. Those cases where age or sex are significantly different between PD and HC are indicated with an asterisk (2 sample t-test for age and Chi-square for sex with 95% confidence).

	Gait		Balance		Voice		Tapping		Multimodal	
	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC
Experiment 1 (all)										
N	610	787	612	803	893	1257	970	1630	597	742
Male/ female	399/ 211*	640/ 147*	401/ 211*	653/ 150*	571/ 322*	1018/ 239*	630/ 340*	1336/ 294*	390/ 207*	607/ 135*
Age (mean±sd)	60.3 ± 8.94*	49.04 ± 10.71*	60.29 ± 8.94*	48.9 ± 10.72*	60.13 ± 8.97*	47.65 ± 10.41*	59.85 ± 9.05*	46.84 ± 10.05*	60.36 ± 8.86*	49.22± 10.78*
UPDRS mean±sd (n)	13.17 ± 7.78 (350)	-	13.14 ± 7.78 (351)	-	13.48 ± 7.93 (566)	-	13.44 ± 7.89 (588)	-	13.15 ± 7.8 (344)	-
UPDRS I mean±sd (n)	5.66 ± 3.64 (361)	-	5.64 ± 3.64 (362)	-	5.64 ± 3.63 (586)	-	5.63 ± 3.63 (608)	-	5.64 ± 3.64 (355)	-
UPDRS II mean±sd (n)	7.59 ± 5.18 (350)	-	7.58 ± 5.18 (351)	-	7.9 ± 5.53 (572)	-	7.86 ± 5.49 (594)	-	7.59 ± 5.21 (344)	-
Experiment 2 (matched)										
N	373	373	376	376	534	534	608	608	361	361
Male/ female	278/ 95	286/ 87	280/ 96	288/ 88	379/ 155	394/ 140	435/ 173	450/ 158	270/ 91	278/ 83
Age (mean±sd)	57.09 ± 9.4	57.09 ± 9.4	57.1 ± 9.4	57.1 ± 9.4	56.54 ± 9.3	56.54 ± 9.3	56.38 ± 9.23	56.38 ± 9.23	57.18 ± 9.36	57.18 ± 9.36
UPDRS mean±sd (n)	14.38 ± 8.48 (206)	-	13.61 ± 8.34 (202)	-	13.76 ± 8.21 (324)	-	13.7 ± 8.08 (349)	-	14.37 ± 8.51 (190)	-
UPDRS I mean±sd (n)	6.16 ± 3.91 (213)	-	5.73 ± 3.85 (207)	-	5.84 ± 3.78 (333)	-	7.8 ± 3.71 (361)	-	5.99 ± 3.82 (198)	-
UPDRS II mean±sd (n)	8.27 ± 5.66 (206)	-	7.95 ± 5.49 (202)	-	7.98 ± 5.6 (328)	-	7.97 ± 5.53 (352)	-	8.45 ± 5.69 (190)	-
Experiment 3 (no comorbidities, matched), experiment 4 (no comorbidities, matched, age controlled), experiment 5 (no comorbidities, matched, sex controlled) and experiment 6 (no comorbidities, matched, controlled)										
N	317	317	320	320	446	446	507	507	306	306
Male/ female	230/ 87	244/ 73	232/ 88	246/ 74	314/ 132	332/ 114	359/ 148	377/ 130	223/ 83	238/ 68
Age (mean±sd)	56.34 ± 9.42	56.34 ± 9.42	56.37 ± 9.41	56.37 ± 9.41	56 ± 9.31	56 ± 9.31	55.71 ± 9.22	55.71 ± 9.22	56.45 ± 9.37	56.45 ± 9.37
UPDRS mean±sd (n)	13.36 ± 7.94 (166)	-	13.53 ± 7.99 (174)	-	13.42 ± 7.63 (275)	-	13.56 ± 7.62 (296)	-	13.5 ± 7.95 (165)	-
UPDRS I mean±sd (n)	5.77 ± 3.71 (172)	-	5.84 ± 3.71 (179)	-	5.56 ± 3.52 (284)	-	5.81 ± 3.54 (304)	-	5.86 ± 3.65 (172)	-
UPDRS II mean±sd (n)	7.65 ± 5.4 (166)	-	7.77 ± 5.44 (174)	-	7.95 ± 5.41 (278)	-	7.85 ± 5.34 (301)	-	7.75 ± 5.38 (165)	-

studies eluded the importance of matching and controlling for these variables [24]–[26], including age, sex [24], [27] or comorbidities which might induce motor (i.e. bradykinesia, tremor or rigidity) and non-motor (i.e. fatigue, restless legs or sleep) symptoms [25]. Other potential data collection biases include small sample sizes [14], [28], inclusion of several recordings per subject [15], [24] or signals of different time lengths [27], which may potentially lead the classifier to detect the idiosyncrasies of each subject rather than specific PD related symptoms, as demonstrated by Neto *et al.* [29]–[31]. In addition, replicability of results is rarely performed in current studies, which may lead to lack of generalizability. Despite the considerable promise for DB in healthcare, these issues limit comparability across studies, hindering interpretation and obstructing translation to the clinic.

Recently, a large dataset of at-home smartphone-based assessments of commonly applied PD tasks including gait, balance, finger tapping and voice evaluations was collected in the mPower study providing a unique resource to examine DB in the study of PD [32], [33]. Indeed, several studies applying machine learning (ML) algorithms have employed this dataset in the study of PD diagnosis, achieving quite different results across studies. Whilst plausible, the impact of the aforementioned confounds on ML-based detection of PD using different at-home digital assessments has not been yet systematically established and has indeed been ignored in many previous studies [15], [24], [27], [34], [35].

Here we systematically explore the influence of accounting for age, sex and comorbidities in the detection of PD in a large at-home dataset. Concretely, we use the mPower dataset to evaluate the ability of common DB task (gait, balance, voice, tapping) for differentiation between PD and HC. In addition, we identify potential DB of Parkinson's disease. With this work, we aim to outline practical suggestions to guide future studies practices and improve comparability across studies.

II. METHODS

A. DATA

Data used in this work were derived from the mPower study [32]. MPower is a mobile application-based study to monitor indicators of PD progression and diagnosis by the

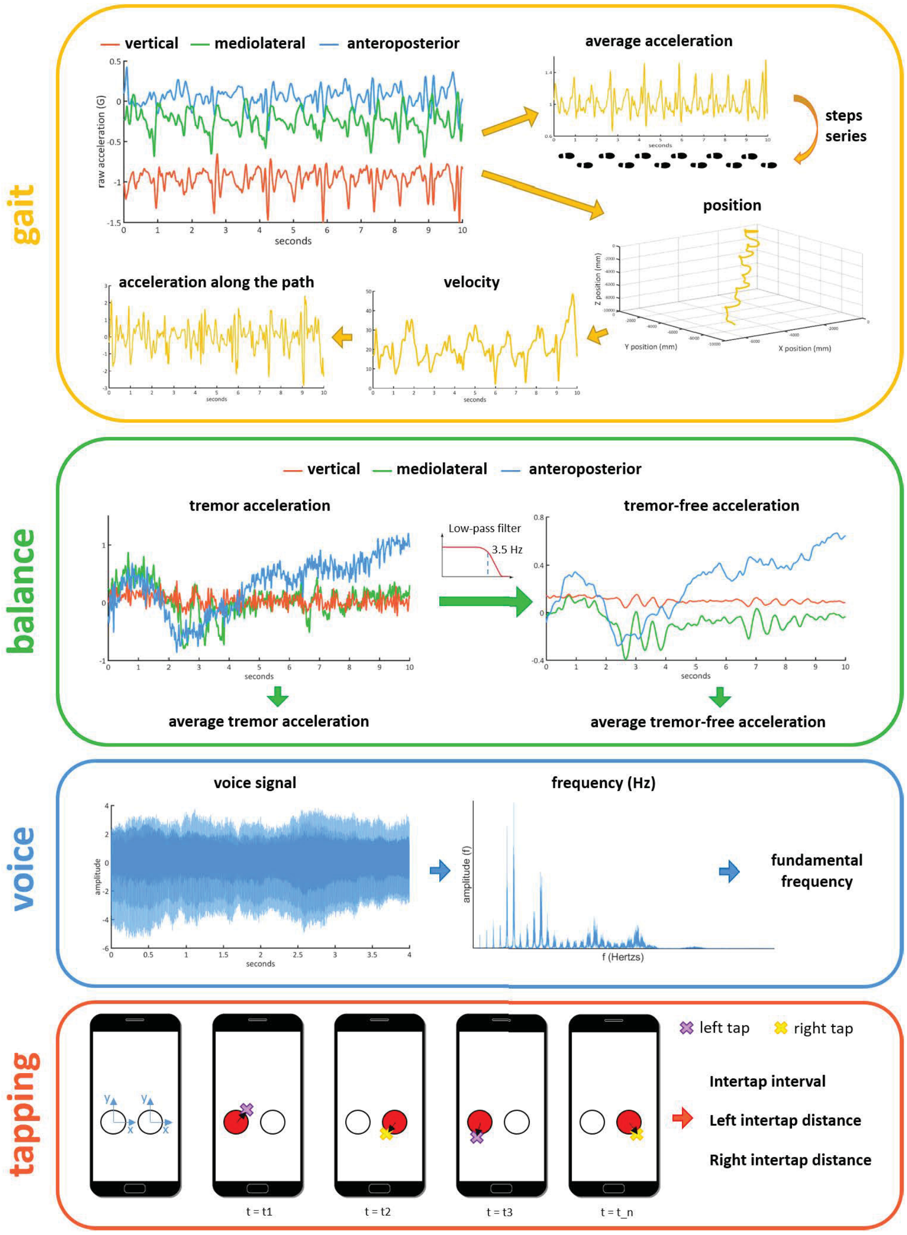


FIGURE 1. Illustration of signal processing and feature extraction based on the raw data for each task.

TABLE 2. List of experiments indicating their corresponding processing steps.

	Exclude comorbidities	Age & sex matching	Control for age	Control for sex	Control for age & sex
E1	-	-	-	-	-
E2	-	x	-	-	-
E3	x	x	-	-	-
E4	x	x	x	-	-
E5	x	x	-	x	-
E6	x	x	x	x	x

E: Experiment

TABLE 3. Balanced accuracy results for CV and holdout datasets and chance level at 95%.

	Experiment 1			Experiment 2			Experiment 3			Experiment 4			Experiment 5			Experiment 6		
	CV	H	C	CV	H	C	CV	H	C	CV	H	C	CV	H	C	CV	H	C
Gait	56.6 (54.3-58.9)	57.1	52.6 (51.1-54.2)	50.3 (47.2-53.6)	54.8	50 (47.1-52.6)	56.5 (53.3-59.7)	55.7	49.9 (46.6-53.4)	56.5 (53.3-59.5)	54.8	50 (46.7-53.3)	56.4 (53.1-59.7)	56.2	50 (46.4-53.6)	56.7 (53.3-59.9)	53.8	50.1 (46.9-53.2)
Balance	61.8 (60.4-63.4)	64.7	50.2 (45.7-54.4)	60.4 (58.6-62.4)	58	49.8 (43.2-56)	60 (57.6-62.3)	59.9	49.9 (43.4-56.6)	60.6 (57.2-63.8)	61.3	50.1 (43.4-56.6)	60.1 (56.5-63.3)	61.3	50.1 (43.4-57.6)	60.2 (57.0-63.6)	59.9	50.2 (42.9-57.1)
Voice	62.5 (61.3-63.6)	60.4	50 (46.5-53.5)	53.9 (51.5-56.2)	59.8	50.1 (44.7-55.3)	56.7 (54.4-58.9)	53	50 (43.6-55.7)	56.9 (54.7-59.1)	58.1	50 (44.6-56.1)	60 (57.7-62.1)	60.1	49.8 (43.9-55.4)	59.2 (57.1-61.2)	59.1	50.2 (43.6-55.7)
Tapping	74.8 (74.4-75.2)	72.9	50 (47-52.9)	66.8 (66-67.6)	66.8	49.9 (45.1-55)	67.9 (67-68.9)	67.2	50.1 (44.1-55.9)	68.8 (67.9-69.8)	66.9	50 (44.4-55)	68.7 (67.6-69.7)	68.9	50.2 (45.3-55.3)	68.8 (67.8-69.8)	68.1	50 (45-55.6)
Multimodal	73.9 (72.4-75.5)	76.9	50 (45.1-54.9)	69.4 (67-71.9)	70	50.1 (44.2-56.7)	69.6 (66.9-72.4)	73.5	50 (43.1-56.9)	69.2 (66.2-71.8)	73	50.2 (43.6-56.4)	68 (65-70,8)	69.1	50 (43.6-57.4)	69.9 (67.2-72.8)	70.6	50 (43.1-56.9)

collection of data in subjects with and without PD. Using this app, subjects were presented with a one-time demographic survey about general demographic topics and health history. Completion of the Movement Disorder Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and the Parkinson's Disease Questionnaire short form (PDQ-8) surveys used for PD assessment was requested at baseline as well as monthly throughout the course of the study. Due to the length of the MDS-UPDRS instrument, subjects were presented only a subset of questions focusing largely on the monitor symptoms of PD [32]. Participants had to select ''true'' or ''false'' to the following question ''Have you been diagnosed by a medical professional with Parkinson Disease?''. According to this answer, they were classified as Parkinson's Disease (PD) or Healthy Control (HC). Subjects who did not answer this question were discarded from further analysis. All subjects were presented with different tasks including gait, balance, voice and tapping, which they could complete up to 3 times per day. Subjects who self-identified as having a professional diagnosis of PD were asked to perform these tasks (1) immediately before taking their medication, (2) after taking their medication and (3) at some other time (Table 8). Subjects who self-identified as not having a diagnosis of PD could complete these tasks at any time during the day. In the gait task, subjects were asked to walk 20 steps in a straight line. In the balance task they were required to stand still for 30 seconds. During the voice activity task, subjects were requested to say 'Aaah' into the microphone for 10 seconds. Finally, during the tapping task participants were instructed to alternatively tap two points on the screen within a 20 seconds interval. We additionally excluded those subjects who gave no information about their age, sex or had inconsistencies in their clinical data (e.g. self-reported healthy controls who answered questions about PD diagnosis or PD medication). Since the mPower dataset is strongly slanted toward young HC (Table 15), we restricted our analysis to those subjects within the age range of 35 to 75 years old. This cleaning step resulted in the exclusion of 40-50% of the data depending on the task. To avoid ''learning effects'' and biases due to several recordings, we only considered the first recording of each

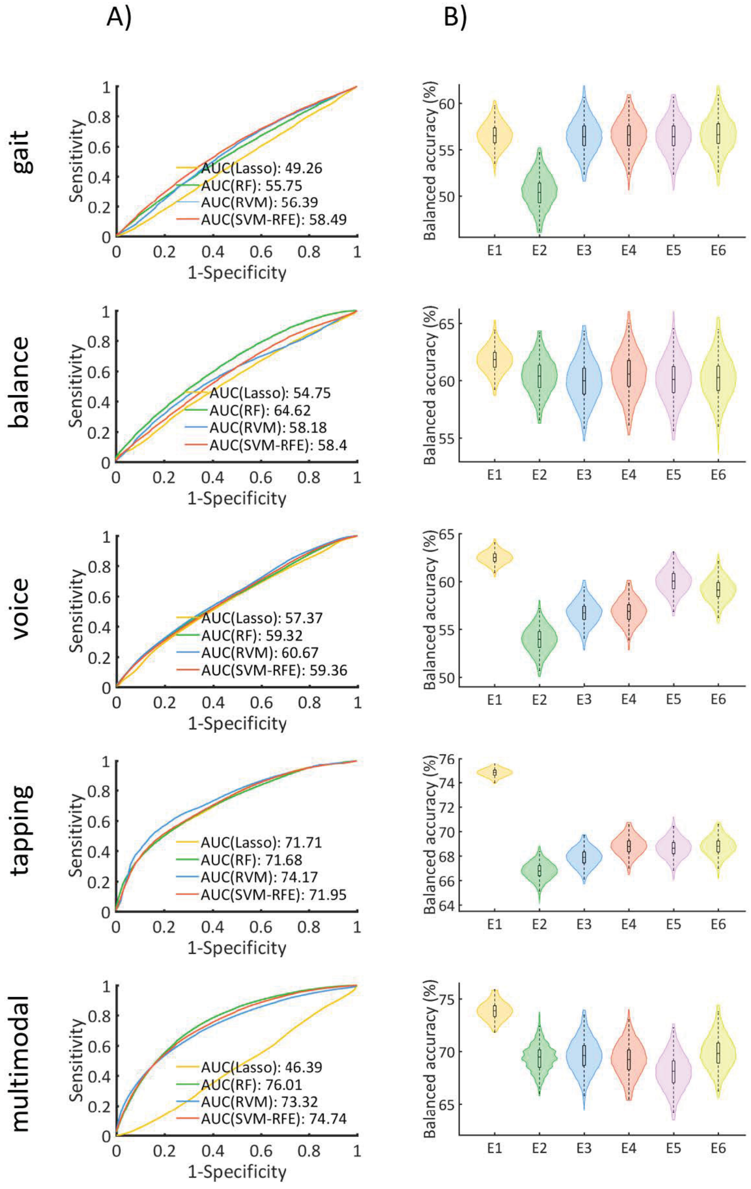


FIGURE 2. A) ROC curves and AUC values for 4 different classifiers for each task, during the main experiment (E3: no comorbidities, matched). B) Balanced accuracy distributions for each task and experiment (E1-E6). E1: all data. E2: age and sex matched. E3: no comorbidities, age and sex matched. E4: no comorbidities, age and sex matched, controlled for age. E5: no comorbidities, age and sex matched, controlled for sex. E6: no comorbidities, age and sex matched, controlled for age and sex.

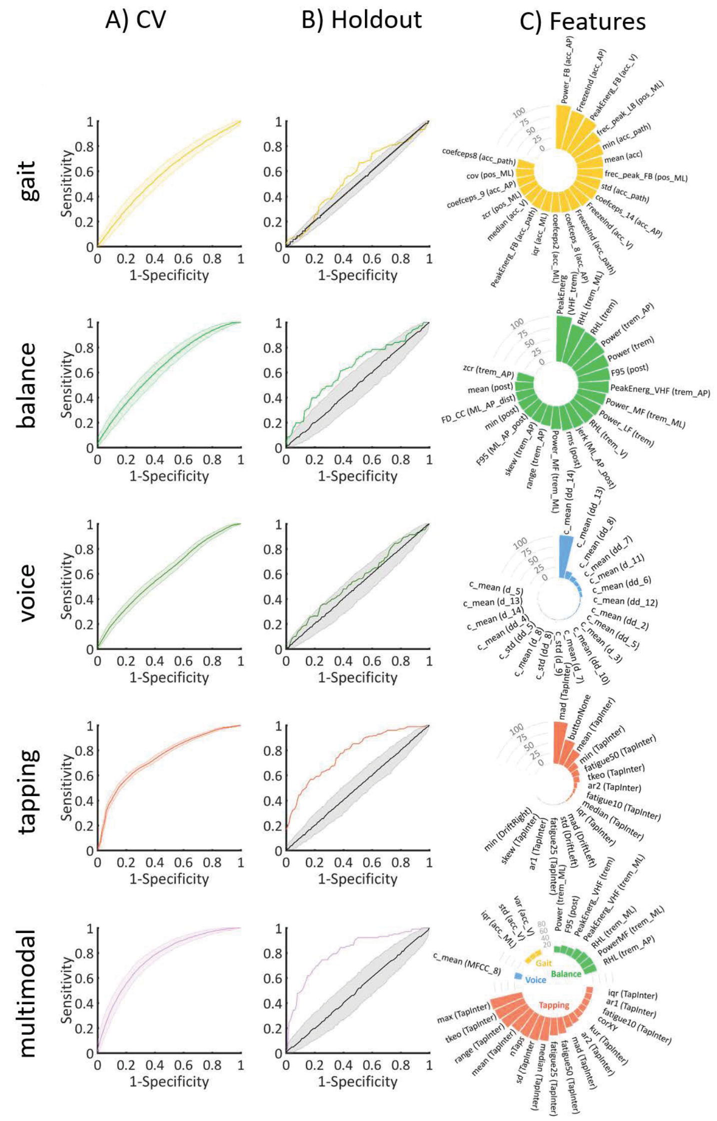


FIGURE 3. A) ROC curves at 95% CI during CV. B) ROC curves at 95% CI during validation of holdout set and at the chance level. C) Scaled average weights of features for each task for the main experiment (E3: no comorbidities, matched). Gait) acc - average acceleration, acc_path – acceleration along path, AP – anteroposterior, FB – freezing band, LB – locomotor band, ML – mediolateral, pos – position, V – vertical, vel – velocity. Balance) trem – tremor, post – postural, dist – distance, LF – low frequency, MF – medium frequency, VHF – very high frequency, RHL – ratio between high and low frequency, F95 –frequency containing 95% of the power spectrum. Voice) c – cepstral coefficient, d – 1st derivative of cepstral coefficient, dd – 2nd derivative of cepstral coefficient. Tapping) TapInter – tap interval. For details on features refer to Appendix A.

TABLE 4. List of gait features.

Feature acronym	Feature description	Signal (acronym)
numSteps	Number of steps during the 10 seconds gait signal.	
MSI	Mean Stride Interval, calculated as the duration of a stride averaged over all strides [58], [64].	
StrideVar	Stride Variability, calculated as the standard deviation divided by the mean stride of the stride interval. Measures consistency and stability [58], [64].	
mean	Mean value of the observations [15], [40].	vertical, anteroposterior, mediolateral and average acceleration (acc_V, acc_AP, acc_ML, acc) vertical, anteroposterior, mediolateral and average position (pos_V, pos_AP, pos_ML, pos) velocity (vel) acceleration along path (acc_path)
min	Minimum value of the observations.	
max	Maximum value of the observations.	
median	Median. Middle value among a dataset [15], [40].	
sd	Standard deviation, calculated as the sum of squares differences between the individual values and the mean. Measures variability [15], [40].	
var	Variance, calculated as the square of the standard deviation. Measures variability.	
range	Range of the observations.	
iqr	Interquartile range, calculated as the difference between 75th and 25th percentiles. Measures dispersion [15], [40].	
rms	Root mean square of the observations.	
cov	Coefficient of variation, calculated as the standard deviation of the signal divided by the mean.	
skew	Skewness. Describes the asymmetry of a signal. A negative value indicates that the distribution is concentrated on the right, while a positive one is concentrated in the left [15], [40], [58].	
kur	Kurtosis. Measures if data is heavy or light-tailed to a normal distribution [40], [58].	
zcr	Zero-crossing rate. Rates sign-changes along a signal [15].	
ApEn	Entropy. Measures uncertainty, ranging from 0-1 where 0 indicates randomness and 1 maximum regularity [15], [58].	
PeakEnerg_LB	Peak of energy in the locomotor band (0.5-3 Hz) [41] , [64].	
frec_peak_LB	Frequency at the peak of energy in the locomotor band (0.5-3 Hz) [41].	
Power_LB	Power of the locomotor band (0.5-3 Hz) [41].	
PeakEnerg_FB	Peak of energy in the freezing band (3-8Hz) [42].	
frec_peak_FB	Frequency at the peak of energy in the freezing band (3-8 Hz).	
Power_FB	Power in the freezing band (3-8 Hz).	
FreezeInd	Freeze Index. Calculated as the ratio between the power in the freezing band (3-8 Hz) and the power in the locomotor band (0.5-3 Hz) [42].	
RatioPower	Sum of the power in the freezing (3-8 Hz) and locomotor band (3-8 Hz) [43]	
ar	Coefficient of a 1st order autoregressive model. An autoregressive model forecasts when there is some correlation between current values and their preceding ones [40].	
coefceps_(1-20)	20 Mel Frequency Cepstral Coefficients. Represent the short-term power spectrum [42]	

subject in the analyses. Further details about data cleaning can be found in Appendix A. Demographic details are shown in Table 1.

B. PRE-PROCESSING

The tri-axial accelerometer integrated in the smartphone records acceleration in the 3 axes (vertical, mediolateral and anteroposterior) during the gait and balance tasks. A 4th order 20 Hz cut-off low-pass Butterworth filter was applied to the 3 accelerometer signals. An additional 3rd order 0.3 Hz cut-off high-pass Butterworth filter was applied to minimize the acceleration variability due to respiration [36]. Signals were then standardized to eliminate the gravity component while maintaining the information from outlier data. According to Pittman *et al.* [24], 30% of the devices were not held in the correct position and therefore, we additionally calculated the average acceleration signal. Several signals were extracted from the gait recordings including the step series, position along the 3 axes calculated by double integration, velocity and acceleration along the path [37] (Figure 1).

Two additional signals were considered for the balance task (Figure 1). Tremor frequency in PD is estimated to fall in the 4-7 Hz band [38], whereas postural acceleration measures (tremor-free) fall in the 0-3.5 Hz interval. To extract tremor-free measures of postural acceleration, we applied a 3.5 Hz cut-off low-pass Butterworth filter [39].

Voice was recorded at a sample rate of 44.1 Kbps. Pre-processing included a downsampling to 25 KHz and a noise reduction using a 2nd order Butterworth filter with a low-pass frequency at 400 Hz. The fundamental frequency signal was calculated using a Hamming window of 20 ms with 50% overlap, and verified with the software Praat (Figure 1). Time, frequency and amplitude series were extracted from the voice signals.

TABLE 5. List of balance features.

Acronym	Description	Signal (acronym)
mean	Mean value of the observations.	vertical, anteroposterior, mediolateral and average tremor acceleration (trem_V, trem_AP, trem_ML, trem) vertical, anteroposterior, mediolateral and average postural acceleration (post_V, post _AP, post _ML, post)
min	Minimum value of the observations.	
max	Maximum value of the observations.	
median	Median value of the observations.	
sd	Standard deviation, calculated as the sum of squares differences between the individual values and the mean. Measures variability.	
var	Variance, calculated as the square of the standard deviation. Measures variability.	
range	Range of the observations.	
iqr	Interquartile range, calculated as the difference between 75th and 25th percentiles. Measures dispersion.	
rms	Root mean square of the observations.	
cov	Coefficient of variation, calculated as the standard deviation of the signal divided by the mean.	
skew	Skewness. Describes the asymmetry of a signal. A negative value indicates that the distribution is concentrated on the right, while a positive one is concentrated in the left.	
kur	Kurtosis. Measures if data is heavy or light-tailed to a normal distribution.	
zcr	Zero-crossing rate. Rates sign-changes along a signal.	
ApEn	Entropy. Measures uncertainty, ranging from 0-1 where 0 indicates randomness and 1 maximum regularity.	
Power_MF	Power of the medium frequency band (4-7Hz) [39].	vertical, anteroposterior, mediolateral and average tremor acceleration (trem_V, trem_AP, trem_ML, trem)
PeakEnerg_VHF	Peak of energy in the very high frequency band (>7Hz)	
frec_peak_HF	Frequency at the peak of energy in the high frequency band (>4Hz) [39].	
Power	Power between 3.5-15Hz	
Power_LF	Power in the low frequency band (0.15-3.5Hz)	
RHL	Ratio between the power between 3.5-15Hz and power between 0.15-3.5Hz [39].	
CFREQ	Centroidal frequency for postural measures. Also known as zero-crossing frequency [36], [39], [44].	anteroposterior, mediolateral and average postural acceleration (post _AP, post _ML, post) mediolateral-anteroposterior average postural acceleration (ML_AP_post)
FREQD	Frequency of dispersion of the power spectrum for postural measures [36], [39], [44].	
jerk	Average jerk. Measures vibration as the rate of change in acceleration. Calculated as the derivative of acceleration with respect to time [36], [39].	vertical, anteroposterior, mediolateral and average postural acceleration (post_V, post _AP, post _ML, post) mediolateral-anteroposterior average postural acceleration (ML_AP_post)
TotalPower	Energy between 0.15-3.5Hz for postural measures [36].	
F50	Frequency containing 50% of the total power for postural measures [36], [39].	
F95	Frequency containing 95% of the total power for postural measures [36], [39].	
MDIST	Represents the average distance from the center to each AP and ML points [39], [44].	mediolateral, anteroposterior and average of mediolateral and anteroposterior distance (ML-dist, AP_dist, ML_AP_dist)
RDIST	Root Mean Square distance from the mean center [44].	
TOTEX	Total excursions is the total length of the path. Calculated as the sum of distances between consecutive points [44].	
MVELO	Mean velocity is the average velocity of the center path, calculated as the TOTEX divided by the time [39], [44].	
MFREQ	The mean frequency is the rotational frequency with a radius equal to the mean distance [36], [44].	
AREA_CC	The 95% confidence circle area is the area of a circle enclosing all points in the AP-ML plane with 95% confidence [36], [44].	average of mediolateral and anteroposterior distance (ML_AP_dist)
AREA_CE	The 95% confidence ellipse area is the area of an ellipse enclosing all points in the AP-ML plane with 95% confidence [36], [39], [44].	
AREA_SW	Sway area calculated as the area enclosing the acceleration path [36], [39], [44].	
FD	The fractal dimension indicates the degree to which a curve fills the enclosed metric space [36], [44].	
FD_CC	Fractal dimension based on the 95% confidence circle area [36], [44].	
FD_CE	Fractal dimension based on the 95% confidence ellipse area [36], [44].	

TABLE 6. List of voice features.

Feature acronym	Feature description	Signal (acronym)
amp	Average amplitude [45].	
shim	Absolute shimmer [15], [26], [45].	
shdb	Shimmer in logarithmic domain [45].	
apq3	3 point amplitude perturbation quotient in percentage [45].	
apq5	5 point amplitude perturbation quotient in percentage [45].	
fm	Frequency modulation [45].	
hnr_mean	Mean of the harmonic to noise ratio, which indicates the amount of noise [15], [26], [45].	
hnr_std	Standard deviation of the harmonic to noise ratio [45].	
rpde	Recurrence period density entropy. Characterizes the deviation from signal periodicity [15], [45].	
DFA	Detrended Fluctuation Analysis, which describes turbulent noise [15], [45].	
mean	Mean value [15], [45].	fundamental frequency (f0), amplitude (amp), Teager Kaiser Energy Operator of the fundamental frequency (tkeo_f0), open quotient (oq), glottis quotient open (gqo), glottis quotient closed (gqc)
sd	Standard deviation [15], [45].	
jitt	Absolute jitter [15], [45].	fundamental frequency (f0), period (T)
jitta	Relative or local jitter [45].	
rap	Relative average perturbation [45].	
ppq5	Perturbation quotient using 5 point (cycles) [45].	
range	Range [45].	fundamental frequency (f0)
tkeo_p25	25th percentile of the Teager-Kaiser Energy Operator [45].	
tkeo_p75	75th percentile of the Teager-Kaiser Energy Operator [45].	
ApEn	Pitch Period Entropy. Quantifies the impaired control of stable pitch during a sustained phonation [15], [45].	
p5	5th percentile [45].	Teager Kaiser Energy Operator of the fundamental frequency (tkeo_f0), open quotient (oq), glottis quotient open (gqo), glottis quotient closed (gqc)
p95	95th percentile [45].	
c_mean	Mean of the Mel Frequency Cepstral Coefficients (MFCCs) coefficients, log-energy of the signal and the first and second derivatives of the MFCCs [15], [26], [45].	log energy (log), 0th order cepstral coefficient (0th), 1-12th Mel Frequency Cepstral Coefficients (MFCC_(1-12), 1-14th deltas (d_(1-14)), 1-14th delta-delta (dd_(1-14))
c_std	Standard deviation of the MFCCs coefficients, log-energy of the signal and the first and second derivatives of the MFCCs [26], [45].	

Tapping recordings consist of the {x,y} screen pixel coordinates and timestamp for each tap on the screen. Both the inter-tapping interval (time) and the {x,y} inter-tap distance series were computed (Figure 1). Further details about preprocessing for each task can be found in Appendix A.

C. FEATURE EXTRACTION

A comprehensive search was conducted in PubMed (https://pubmed.ncbi.nlm.nih.gov/) with the following search terms ((Parkinson's disease) AND (walking OR gait OR balance OR voice OR tapping) AND (wearables OR smartphones)) to identify features commonly applied for each task and corresponding signals generated. Based on the results of this search, 423, 183, 124 and 43 features were identified and computed using Matlab R2017a from gait [40], [42], [43], balance [7], [36], [39], [44], voice [25], [26], [45] and tapping data [15], [32], [46], respectively (Table 4-Table 7).

D. MACHINE LEARNING ALGORITHMS

As a different ML algorithm may provide the best performance for a given task, we evaluated four commonly applied algorithms for differentiation between PD and HC:

1) Least Absolute Shrinkage and Selection Operator (LASSO) is a linear method commonly used to deal with high-dimensional data. LASSO applies a regularization process, where it penalizes the coefficients of the regression variables shrinking some of them to zero. During the feature selection process, those variables with non-zero coefficients are selected to be part of the model [47]. LASSO performs well when dealing with linearly separable data and avoiding overfitting.
2) Random Forest (RF) uses an ensemble of decision trees, where each individual tree outputs the classes. The predicted class is decided based on majority vote. Each tree is built based on a bootstrap training set that normally represents two thirds of the total cohort. The left out data is used to get an unbiased estimate of the classification error and get estimates of feature importance. RF runs efficiently in large datasets and deals very well with data with complicated relationships [48].
3) A Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel with Recursive Feature Elimination (SVM-RFE). An SVM is a linear method whose aim is to find the optimal hyperplane that separates

TABLE 7. List of tapping features.

Feature acronym	Feature description	Signal (acronym)
nTaps	Number of taps [46], [65].	
buttonNone	Frequency of tapping outside the button [46], [65].	
corXY	Correlation between X and Y touchscreen coordinates [46], [65].	
mean	Mean value of the observations [15], [46], [65].	Intertap interval (TapInter), Leftdrift (DriftLeft), Right drift (DriftRight)
min	Minimum value of the observations [15], [46], [65].	
max	Maximum value of the observations [15], [46], [65].	
median	Median value of the observations [15], [46], [65].	
mad	Median absolute deviation [46], [65].	
sd	Standard deviation [15], [46], [65].	
range	Range of the observations [15], [46], [65].	
iqr	Interquartile range [46], [65].	
cov	Coefficient of variation [15], [46], [65].	
skew	Skewness [46], [65].	
kur	Kurtosis [46], [65].	
tkeo	Teager-Kaiser Energy Operator. Measures energy variation [15], [46], [65].	Intertap interval (TapInter)
dfa	Detrended Fluctuation Analysis. Measures changes in the signal [15], [46], [65].	
ar1	Coefficient of an autoregressive model at lag 1. Indicates associations between intertap intervals [15], [46], [65].	
ar2	Coefficient of an autoregressive model at lag 2. Indicates associations between intertap intervals [15], [46], [65].	
fatigue10	Increase in the mean intertap interval from the first 10% to the last 10% taps [15], [46], [65].	
fatigue25	Increase in the mean intertap interval from the first 25% to the last 25% taps [15], [46], [65].	
fatigue50	Increase in the mean intertap interval from the first 50% to the last 50 % taps [15], [46], [65].	

between classes. When data is linearly non-separable, it may be transformed to a higher dimensional space using a non-linear transformation function that spreads the data apart such that a linear hyperplane can be found in that space. Here, we used a radial basis kernel function. RFE is a feature selection method that ranks features according to importance, improving both efficiency and accuracy of the classification model. This model is known to remove effectively non-relevant features and achieve high classification performance [49].

4) Relevance Vector Machine (RVM), which follows the same principles of SVM but provides probabilistic classification. The Bayesian formulation prevents from tuning the hyper-parameters of the SVM. Nonetheless, RVMs use an expectation maximization (EM)-like learning that can lead to local minima unlike the standard sequential optimization (SMO)-based algorithms used by SVMs, that guarantee to find a global optima [50].

E. FRAMEWORK

The following six experiments were performed to address the questions on the impact of age, sex and comorbidities that may influence task performance on the classification accuracy for each task and on the combination of all tasks for differentiation between PD and HC (Table 2):

1) Experiment 1 (E1: all) includes all subjects only restricting the age range (35-75 years old).
2) Experiment 2 (E2: matched) includes subjects after an age and sex matching between PD and HC, where we strictly match one HC for each PD subject with the same age and where possible with the same sex.
3) Experiment 3 (E3: no comorbidities, matched) excludes all comorbidities that may affect task performance (see Appendix A) and strictly matches for age and where possible sex on the remaining subjects.
4) Experiments 4-6 (E4-6): Three additional experiments assess if controlling for age and sex impacts the results. These experiments exclude comorbidities, match for age and sex and control for age and/or sex applying multiple regression. For this, age and gender were included as covariates in a multiple regressions using the features for each modality as dependent variables. The estimated beta coefficients for each covariate were used to regress out the estimated effects of age and sex on the respective feature. The resulting residuals for each feature were used for subsequent classification. Experiment 4 (E4): no comorbidities, matched, controlled for age; Experiment 5 (E5): no comorbidities, matched, controlled for sex; Experiment 6 (E6): no comorbidities, matched, controlled for age and sex.

As the performance obtained after removing comorbidities and matching for age and sex (E3) provides a relatively unbiased estimate for differentiation between PD

TABLE 8. Medication status at the time of performing the tasks.

			PD/Total	Before	After	Another	No Med	Empty
E1	Gait	All	653/2711	154	166	259	63	11
		CV	436/2711	99	120	169	39	9
		Holdout	217/903	55	46	90	24	2
	Balance	All	655/2747	156	166	257	64	12
		CV	437/1832	103	99	186	41	8
		Holdout	218/915	53	67	71	23	4
	Voice	All	965/4799	222	229	396	94	24
		CV	644/3200	140	159	265	62	18
		Holdout	321/1599	82	70	131	32	6
	Tapping	All	1054/6221	237	237	446	106	28
		CV	703/4148	156	157	299	72	19
		Holdout	351/2073	81	80	147	34	9
E2	Gait	All	373/746	91	101	135	39	7
		CV	249/498	62	70	86	28	3
		Holdout	124/248	29	31	49	11	4
	Balance	All	376/752	99	96	139	36	6
		CV	251/502	68	60	97	21	5
		Holdout	125/250	31	36	42	15	1
	Voice	All	534/1068	135	131	205	48	15
		CV	356/712	94	89	127	34	12
		Holdout	178/356	41	42	78	14	3
	Tapping	All	608/1216	134	135	262	59	18
		CV	406/812	83	92	182	39	10
		Holdout	202/404	51	43	80	20	8
E3	Gait	All	317/634	82	84	117	28	6
		CV	212/424	54	53	82	19	4
		Holdout	105/210	28	31	35	9	2
	Balance	All	320/640	76	89	116	32	7
		CV	214/428	48	53	87	22	4
		Holdout	106/212	28	36	29	10	3
	Voice	All	446/892	112	103	190	34	7
		CV	298/596	75	71	125	21	6
		Holdout	148/296	37	32	65	13	9
	Tapping	All	507/1014	124	112	211	44	16
		CV	338/676	80	70	147	30	11
		Holdout	169/338	44	42	64	14	5
E4-6	Gait	All	317/634	82	84	117	28	6
		CV	212/424	54	53	82	19	4
		Holdout	105/210	28	31	35	9	2
	Balance	All	320/640	76	89	116	32	7
		CV	214/428	48	53	87	22	4
		Holdout	106/212	28	36	29	10	3
	Voice	All	446/892	112	103	190	34	7
		CV	298/596	75	71	125	21	6
		Holdout	148/296	37	32	65	13	1
	Tapping	All	507/1014	124	112	211	44	16
		CV	338/676	80	70	147	30	11
		Holdout	169/338	44	42	64	14	5

Before - "Immediately before taking their medication", After - "After taking their medication (when they are feeling at their best)", Another - "At some other time", No Med - "I don't take Parkinson's medication", Empty - question unanswered

and HC, these results were used for selection of the best performing ML algorithm for each task and interpretation of the main outcomes throughout this work. Demographic and clinical information for each experiment are provided in Table 1.

Additionally, to compare the performance of our analyses to those in the literature, we performed an analysis including all data without restricting age range (Table 15) and an analysis including all data and both age and sex as features.

F. MODEL PERFORMANCE

Data leakage occurs when information of the holdout test set leaks into the dataset used to build the model, leading to incorrect or overoptimistic predictions. Therefore, in every experiment and task, data was initially split into 2/3 of data to build the predictive model and 1/3 of holdout data to validate this model. To build the model, we performed 1000 repetitions of 10-fold cross-validation (CV) in the 2/3 of the data for each classifier to avoid data leakage and increase robustness. The parameter Lambda of the LASSO model was set to 1 and

TABLE 9. Cross-validation classification performances for each of the tasks (gait, balance, voice, tapping and multimodal features) for four different classifiers.

	% (95% CI)	Gait	Balance	Voice	Tapping	Multimodal
LASSO	Balanced Accuracy	50.14 (48; 52.12)	53.49 (51.4; 55.49)	55.38 (52.69; 57.72)	64.29 (63.02; 65.61)	48.35 (44.85; 51.96)
	Sensitivity	71.07 (65.33; 74.76)	47.39 (44.63; 50.47)	54.55 (48.66; 59.06)	64.89 (63.02; 66.72)	53.37 (37.75; 68.14)
	Specificity	29.22 (25; 33.96)	59.58 (56.31; 62.38)	56.2 (52.85; 60.4)	63.7 (61.54; 65.83)	43.33 (29.66; 58.33)
	PPV	50.1 (48.55; 51.51)	53.97 (51.6; 56.23)	55.46 (52.82; 57.69)	64.14 (62.82; 65.58)	48.45 (45.16; 51.76)
	NPV	50.28 (46.73; 53.84)	53.11 (51.25; 54.99)	55.32 (52.59; 57.81)	64.47 (63.11; 65.92)	48.14 (43.96; 52.37)
	AUC	49.26 (45.95; 52.36)	54.75 (51.84; 57.23)	57.37 (54.83; 59.27)	71.71 (70.2; 73.05)	46.39 (42.05; 50.73)
RF	Balanced Accuracy	54.24 (51.42; 56.84)	59.95 (57.59; 62.27)	56.19 (53.27; 59.4)	65.36 (64.05; 66.64)	69.59 (66.91; 72.43)
	Sensitivity	52.43 (48.59; 55.9)	60.28 (57.01; 63.79)	54.84 (50.84; 59.06)	63.05 (61.24; 64.79)	69.57 (65.69; 73.53)
	Specificity	56.04 (52.36; 59.67)	59.63 (56.31; 62.85)	57.53 (53.69; 61.91)	67.67 (65.68; 69.68)	69.61 (65.69; 73.53)
	PPV	54.4 (51.47; 57.09)	59.9 (57.54; 62.23)	56.37 (53.33; 59.6)	66.11 (64.67; 67.52)	69.61 (66.67; 72.86)
	NPV	54.09 (51.38; 56.67)	60.03 (57.71; 62.51)	56.03 (53.19; 59.12)	64.68 (63.32; 65.95)	69.6 (66.67; 72.8)
	AUC	55.75 (51.73; 59.28)	64.62 (61.72; 67.51)	59.32 (56.71; 62.11)	71.68 (70.23; 73.04)	76.01 (73.79; 78.19)
RVM	Balanced Accuracy	55.42 (53.18; 57.67)	57.07 (54.91; 59)	56.7 (54.36; 58.89)	67.89 (67.01; 68.86)	67.02 (63.48; 70.59)
	Sensitivity	55.54 (52.36; 58.73)	54.05 (51.4; 57.01)	57.74 (54.7; 60.74)	64.34 (63.17; 65.39)	65.43 (60.29; 70.1)
	Specificity	55.31 (52.12; 58.26)	60.09 (57.24; 62.85)	55.66 (52.35; 58.89)	71.43 (70.12; 72.78)	68.6 (63.73; 73.78)
	PPV	55.42 (53.22; 57.62)	57.54 (55.23; 59.69)	56.57 (54.26; 58.63)	69.25 (68.14; 70.42)	67.6 (63.9; 71.69)
	NPV	55.44 (53.14; 57.72)	56.68 (54.59; 58.52)	56.85 (54.47; 59.03)	66.7 (65.8; 67.55)	66.51 (62.74; 70.07)
	AUC	56.39 (53.55; 59.44)	58.18 (55.27; 60.9)	60.67 (58.87; 62.4)	74.17 (73.27; 75.01)	73.32 (70.19; 76.54)
SVM-RFE	Balanced Accuracy	56.5 (53.3; 59.7)	56.19 (52.57; 59.6)	56 (53.2; 58.6)	65.83 (63.76; 67.97)	68.36 (65.47; 71.41)
	Sensitivity	56.56 (53.38; 59.9)	56.17 (52.53; 59.63)	55.56 (53.02; 58.16)	67.61 (65; 70.26)	69.23 (66; 72.63)
	Specificity	56.38 (53.14; 59.6)	56.21 (52.52; 59.95)	56.44 (53.44; 59.21)	64.05 (62.24; 65.94)	67.48 (64.52; 70.79)
	PPV	55.81 (50.94; 60.85)	56.3 (50; 62.62)	59.63 (55.37; 63.59)	60.02 (57.1; 63.02)	65.92 (61.77; 70.59)
	NPV	57.11 (52.36; 61.79)	56.06 (50.47; 62.15)	52.29 (47.65; 56.54)	71.23 (67.46; 74.56)	70.69 (66.67; 74.51)
	AUC	58.49 (55.55; 61.57)	58.4 (54.35; 62.1)	59.36 (56.3; 61.85)	71.95 (70.01; 73.75)	74.74 (72.29; 77.23)

AUC – Area Under the Curve, NPV – Negative Predictive Values, PPV – Positive Predictive Values, RF – Random Forest, RVM – Relevance Vector Machine, SVM-RFE – Support Vector Machine-Recursive Feature Elimination

the number of trees for RF to 100. A nested cross-validation was implemented to tune the parameters of the SVM-RFE classifier. The procedure consists of an inner CV to select the best parameters of the model following a grid search for the regularization constant (C) ranging from 2^{-7} to 2^{7} and for gamma (γ) ranging from 2^{-4} to 2^{4} for the SVM. Then, the outer loop is used to assess the model selected in the inner CV. Extensive parameter optimization was applied only on SVM-RFE classifier, given that the other algorithms have already embedded optimization and that 1000 repetitions of 10-fold cross-validation and multiple experiments would have taken based on the estimated from a single run each at least several months on the high-throughput cluster available to us. For each model, we report the following measures of predictive performance: balanced accuracy (BA), sensitivity, specificity, positive (PPV) and negative predictive value (NPV), mean receiver operating characteristic (ROC) curves with 95% confidence intervals and area under the curve (AUC). Comparisons between models are based on BA.

Once the best predictive model with the highest cross-validation BA was identified using the CV dataset, it was validated using the holdout dataset, reporting the

TABLE 10. Classification performance for the gait task.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	56.56 (54.34; 58.85)	57.11	52.63 (51.12; 54.21)	50.33 (47.19; 53.62)	54.84	50.03 (47.13; 52.58)	56.5 (53.3; 59.7)	55.71	49.92 (46.63; 53.36)
Sensitivity (%)	52.63 (49.52; 55.89)	39.41	31.41 (26.11; 36.45)	50.34 (47.15; 53.75)	50	45.21 (38.71; 50.81)	56.56 (53.38; 59.9)	56.19	50.33 (43.81; 57.14)
Specificity (%)	60.49 (59.12; 62.02)	74.81	73.86 (71.97; 76.13)	50.33 (47.15; 53.54)	59.68	54.85 (54.35; 55.56)	56.38 (53.14; 59.6)	55.24	49.52 (49.45; 49.58)
PPV (%)	38.22 (34.64; 42.02)	54.79	50 (50; 50)	49.68 (45.38; 54.22)	55.36	50 (50; 50)	55.81 (50.94; 60.85)	55.66	50 (50; 50)
NPV (%)	73.32 (70.1; 76.19)	61.44	56.35 (52.98; 59.56)	50.98 (46.19; 55.82)	54.41	50.05 (44.12; 55.15)	57.11 (52.36; 61.79)	55.77	49.85 (43.27; 56.73)
AUC (%)	59.45 (57.41; 61.38)	59.88	50.03 (44.42; 55.17)	50.98 (47.84; 54.13)	56.5	50.07 (42.9; 56.99)	58.49 (55.55; 61.57)	55.85	49.95 (41.48; 57.74)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	56.48 (53.3; 59.46)	54.76	49.97 (46.7; 53.31)	56.4 (53.07; 59.68)	56.19	50 (46.43; 53.57)	56.65 (53.3; 59.91)	53.81	50.05 (46.85; 53.19)
Sensitivity (%)	56.56 (53.24; 59.66)	54.29	49.46 (42.86; 56.19)	56.46 (53.08; 59.95)	56.19	49.99 (42.86; 57.14)	56.71 (53.32; 60.05)	51.43	47.71 (40.95; 54.29)
Specificity (%)	56.39 (53.16; 59.43)	55.24	50.48 (50.42; 50.55)	56.35 (52.93; 59.62)	56.19	50 (50; 50)	56.58 (53.24; 59.91)	56.19	52.39 (52.1; 52.75)
PPV (%)	55.82 (51.42; 60.38)	54.81	50 (50; 50)	56 (50.94; 60.38)	56.19	50 (50; 50)	56.12 (51.42; 60.38)	54	50 (50; 50)
NPV (%)	57.12 (52.36; 61.79)	54.72	49.94 (43.4; 56.6)	56.82 (52.36; 61.32)	56.19	49.99 (42.86; 57.14)	57.15 (52.83; 61.79)	53.64	50.08 (43.64; 56.36)
AUC (%)	58.51 (55.51; 61.3)	56.25	49.99 (41.91; 58.25)	58.2 (54.83; 61.24)	56.1	50.02 (42.65; 58.05)	58.33 (55.25; 61.52)	55.99	50.07 (42.16; 57.99)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

aforementioned performance metrics. In addition, to test whether the BA of the predictive model is higher than chance level (0.5 for binary classification), we ran 1000 permutations randomly permuting the predicted classes, reporting BA at 95% confidence intervals.

III. RESULTS

A. CLASSIFIER SELECTION AND RESULTS FOR THE CV DATASET

Four different classifiers (random forest: RF, Least Absolute Shrinkage and Selection Operator: LASSO, support vector machine: SVM, relevance vector machine: RVM-RFE) were applied to each of the four tasks and their combination during the main experiment (E3: no comorbidities, matched for age and sex). Table 9 provides detailed information on the classification performance for each ML algorithm and each task. The ROC curves and corresponding AUC values for the four classifiers for each of the tasks during the cross-validation (CV) step are displayed in Figure 2A. RF, RVM and SVM-RFE performed similarly across all tasks, whereas LASSO was the classifier performing the poorest. Best performance was achieved on the combination of all tasks using RF (balanced accuracy (BA)): 69.6%), followed by tapping using RVM (BA: 67.9%), balance using RF (BA: 60%), voice using RVM (BA: 56.7%) and gait using SVM-RFE (BA: 56.5%).

B. COMPARISON OF EXPERIMENTS IN THE CROSS-VALIDATION SETTING

ML algorithms performing best for each task in the main experiment (E3: no comorbidities, matched for age and sex) were applied to corresponding task data of the other five

TABLE 11. Classification performance for the balance task.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	61.82 (60.41; 63.4)	64.73	50.15 (45.68; 54.38)	60.42 (58.57; 62.35)	58	49.81 (43.2; 56)	59.95 (57.59; 62.27)	59.91	49.93 (43.4; 56.6)
Sensitivity (%)	44.96 (42.89; 47.43)	50.25	33.72 (28.43; 38.73)	57.72 (54.78; 60.36)	58	49.81 (43.2; 56)	60.28 (57.01; 63.79)	65.57	55.59 (49.06; 62.26)
Specificity (%)	78.68 (76.96; 80.41)	79.21	66.59 (62.55; 70.41)	63.11 (60.36; 66.14)	58	49.81 (44; 56)	59.63 (56.31; 62.85)	54.25	44.27 (37.74; 50.94)
PPV (%)	61.63 (59.37; 64.05)	64.87	43.54 (36.94; 49.69)	61.02 (58.98; 63.14)	58	49.81 (43.31; 56.1)	59.9 (57.54; 62.23)	58.9	49.94 (44.07; 55.93)
NPV (%)	65.26 (64.21; 66.46)	67.57	56.8 (53.5; 59.94)	59.89 (58.1; 61.72)	58	49.81 (43.31; 56.1)	60.03 (57.71; 62.51)	61.17	49.92 (42.71; 57.61)
AUC (%)	67.02 (65.38; 68.73)	70.45	50.15 (44.85; 51.19)	65.15 (62.85; 67.43)	61.02	49.86 (42.16; 57.21)	64.61 (61.72; 67.51)	63.09	49.99 (42.18; 57.44)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	60.58 (57.24; 63.79)	61.32	50.11 (43.4; 56.6)	60.12 (56.54; 63.32)	61.32	50.08 (43.4; 57.55)	60.24 (57.01; 63.55)	59.91	50.2 (42.93; 57.08)
Sensitivity (%)	60.67 (55.84; 65.42)	66.04	54.82 (48.11; 61.32)	59.72 (54.67; 64.02)	61.32	50.08 (43.4; 57.55)	60.06 (55.14; 64.95)	58.49	48.78 (41.51; 55.66)
Specificity (%)	60.49 (56.08; 64.95)	56.6	45.39 (38.68; 51.89)	60.5 (56.08; 64.95)	61.32	50.08 (43.4; 57.55)	60.42 (56.08; 64.95)	61.32	51.61 (44.34; 58.49)
PPV (%)	60.57 (57.11; 63.92)	60.34	50.1 (43.97; 56.03)	60.2 (56.63; 63.5)	61.32	50.08 (43.4; 57.55)	60.29 (57.05; 63.57)	60.19	50.2 (42.72; 57.28)
NPV (%)	60.61 (57.11; 63.98)	62.5	50.12 (42.71; 57.29)	60.04 (56.49; 63.33)	61.32	50.08 (43.4; 57.55)	60.22 (56.81; 63.7)	59.63	50.19 (43.12; 56.88)
AUC (%)	65.45 (62.32; 68.41)	63.59	50.08 (41.98; 56.94)	64.96 (61.84; 67.77)	62.76	50.06 (42.24; 58.54)	65.24 (62.32; 68.02)	62.65	50.28 (42.04; 58.05)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

experiments (E1: all subjects, E2: matched for age and sex, E4-6: same as E3 but additionally regressing out the effects of age and/or sex). Classification performance for each task and experiment during the CV and over holdout sets is summarized in Table 3 and Table 10-Table 14. BA distributions for each experiment and task during the CV are displayed in Figure 2B.

In the CV, E1 (all data) resulted in the highest but modest BA for all tasks (gait: 56.6%; balance: 61.8%; voice: 62.5%; tapping: 74.8; multimodal combining all four tasks: 73.9%). Removal of comorbidities in E3 had a marginal effect on BA as compared to E2 (matched for age and sex) with increased BA for gait (E2: 50.3%; E3: 56.5%), voice (E2: 53.9%; E3: 56.7%) and tapping (E2: 66.8%; E3: 67.9%) but lower BA for balance (E2: 60.4%; E3: 60.0%). After additionally regressing out the effects of age and/or sex (E4-E6) the change in the BA was negligible for all tasks ($< 1\%$) except for voice when regressing out sex (E3: 56.7%; E5: 60%) and both age and sex (E3: 56.7%; E6: 59.2%) (Table 3, Tables 10–14).

Analyses including all data without trimming for age range led to the highest accuracy of 74.4% using tapping data, followed by 72.7% for the multimodal case and 58%, 52.9% and 51% for balance, voice and gait data respectively. In all cases specificity was close to 100% whereas sensitivity was exceedingly low (Table 16-Table 20). When including both age and sex as additional features, accuracy increased to 80.8% for tapping data, 75.3% for the multimodal case and 73.1%, 69% and 57% for voice, balance and gait data respectively with high specificities and low sensitivities.

TABLE 12. Classification performance for the voice task.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	62.49 (61.33; 63.62)	60.41	49.99 (46.46; 53.5)	53.94 (51.48; 56.18)	59.83	50.05 (44.66; 55.34)	56.7 (54.36; 58.89)	53.04	49.98 (43.58; 55.74)
Sensitivity (%)	46.54 (44.63; 48.49)	44.44	32.25 (28.11; 36.36)	50.43 (47.05; 53.65)	56.74	46.96 (41.57; 52.25)	57.74 (54.7; 60.74)	52.03	48.97 (42.57; 54.73)
Specificity (%)	78.43 (77.15; 79.71)	76.37	67.73 (64.8; 70.64)	57.45 (53.93; 60.67)	62.92	53.14 (47.75; 58.43)	55.66 (52.35; 58.89)	54.05	50.99 (44.6; 56.76)
PPV (%)	60.55 (58.79; 62.19)	57.14	41.47 (36.15; 46.75)	54.24 (51.62; 56.67)	60.48	50.05 (44.31; 55.69)	56.57 (54.26; 58.63)	53.1	49.98 (43.45; 55.86)
NPV (%)	67.35 (66.5; 68.19)	65.98	58.51 (55.98; 61.03)	53.68 (51.36; 55.76)	59.26	50.04 (44.97; 55.03)	56.85 (54.47; 50.03)	52.98	49.98 (43.71; 55.63)
AUC (%)	68.99 (68.14; 69.79)	66.95	49.93 (45.49; 54.03)	55.5 (53.36; 57.6)	62.48	50.14 (44.18; 56.09)	60.67 (58.87; 62.4)	54.99	50.01 (43.02; 56.65)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	56.85 54.7; 59.06)	58.11	49.98 (44.6; 56.08)	59.99 (57.72; 62.08)	60.14	49.82 (43.92; 55.41)	59.15 (57.05; 61.24)	59.12	50.15 (43.58; 55.74)
Sensitivity (%)	56.64 (53.69; 59.56)	53.38	45.25 (39.87; 51.35)	59.55 (56.38; 62.42)	53.38	43.06 (37.16; 48.65)	58.83 (56.04; 61.41)	58.78	49.82 (43.24; 55.41)
Specificity (%)	57.06 (54.03; 60.07)	62.84	54.71 (49.32; 60.81)	60.43 (57.38; 63.42)	66.89	56.58 (50.68; 62.16)	59.46 (56.71; 62.42)	59.46	50.49 (43.92; 56.08)
PPV (%)	56.88 (54.59; 59.25)	58.96	49.98 (44.03; 56.72)	60.08 (57.76; 62.2)	61.72	49.79 (42.97; 56.25)	59.21 (57; 61.36)	59.18	50.15 (43.54; 55.78)
NPV (%)	56.82 (54.64; 59.06)	57.41	49.98 (45.06; 55.56)	59.91 (57.69; 62)	58.93	49.84 (44.64; 54.76)	59.09 (57.02; 61.23)	59.06	50.15 (43.62; 55.71)
AUC (%)	58.44 (56.37; 60.38)	59.13	50.03 (43.7; 56.47)	63.3 (61.25; 65.05)	61.64	49.85 (43.28; 56.58)	61.72 (59.65; 63.53)	63.26	50.25 (43.32; 57.15)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

C. RESULTS FOR THE HOLDOUT DATASET

Best performing classifiers trained on the 2/3 of the initial dataset used for cross-validation were applied to the 1/3 holdout dataset. Results for the holdout dataset were highly similar to the CV results (Table 3, Tables 10–14). All results are summarized in Figure 3 and Table 3. The multimodal combination of all tasks resulted in the best performance for differentiation of PD and HC in the holdout cohort (BA: 73.5%) followed by the tapping features (67.2%). Voice features achieved the lowest BA of 53% followed by gait (55.7%) and balance (59.9%) features (Table 3). For the base experiment E3, the difference in BA between CV and holdout sets was less than 1% for all tasks except for a 3.7% reduction in BA for voice data and a 3.9% increase for the multimodal feature combination. Exclusion of comorbidities resulted in only minor changes for gait, balance and tapping (<2%) with a 6.8% drop only observed using voice data and a 3.5% increase for the multimodal case. BA performance for all tasks increased by 1.4% (gait) to 7.4% (voice) for all tasks when using the dataset only restricting the age range (E1) as compared to E3. No systematic effects of additionally controlling for age and/or sex prior to classification (E4-E6) were observed with BA changes being small and inconsistent across tasks and experiments.

Analyses including all data without trimming for age range reached the highest accuracy in the holdout set of 73.3% using multimodal features, followed by 71.1% for the tapping task and 55.8%, 52.6% and 51.6% for balance, voice and gait data respectively (Table 16-Table 20). When including both age and sex as additional features, accuracy in the holdout data

TABLE 13. Classification performance for the tapping task.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	74.81 (74.41; 75.23)	72.9	49.99 (46.98; 52.9)	66.78 (65.95; 67.55)	66.83	49.86 (45.05; 54.95)	67.89 (67.01; 68.86)	67.16	50.09 (44.08; 55.92)
Sensitivity (%)	61.09 (60.36; 61.75)	57.59	28.86 (25.08; 32.51)	63.77 (62.56; 64.9)	59.9	42.93 (38.12; 48.02)	64.34 (63.17; 65.39)	68.05	50.98 (44.97; 56.81)
Specificity (%)	88.52 (88.13; 88.91)	88.21	71.13 (68.88; 73.3)	69.8 (68.6; 70.94)	73.76	56.8 (51.98; 61.88)	71.43 (70.12; 72.78)	66.27	49.21 (43.2; 55.03)
PPV (%)	76.01 (75.38; 76.71)	74.4	37.29 (32.4; 42)	67.86 (66.88; 68.74)	69.54	49.84 (44.25; 55.75)	69.25 (68.14; 70.42)	66.86	50.09 (44.19; 55.81)
NPV (%)	79.26 (78.95; 79.56)	77.76	62.7 (60.71; 64.61)	65.83 (65.01; 66.55)	64.78	49.88 (45.65; 54.35)	66.7 (65.8; 67.55)	67.47	50.1 (43.98; 56.02)
AUC (%)	83.48 (83.14; 83.77)	84.42	50.01 (46.37; 53.95)	73.36 (72.56; 73.99)	74.97	49.88 (44.38; 55.73)	74.17 (73.27; 75.01)	77.51	50.13 (43.63; 56.88)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	68.8 (67.9; 69.75)	66.86	49.98 (44.38; 55.03)	68.66 (67.6; 69.68)	68.93	50.23 (45.27; 55.33)	68.8 (67.75; 69.75)	68.05	49.97 (44.97; 55.62)
Sensitivity (%)	65.45 (64.35; 66.86)	66.27	49.39 (43.79; 54.44)	65.86 (64.65; 67.01)	68.64	49.94 (44.97; 55.03)	65.6 (64.35; 66.86)	67.46	49.38 (44.38; 55.03)
Specificity (%)	72.16 (70.71; 73.52)	67.46	50.57 (44.97; 55.62)	71.46 (69.97; 72.93)	69.23	50.53 (45.56; 55.62)	72 (70.56; 73.37)	68.64	50.56 (45.56; 56.21)
PPV (%)	70.16 (69.03; 71.31)	67.07	49.98 (44.31; 55.09)	69.78 (68.53; 71.02)	69.05	50.23 (45.24; 55.36)	70.09 (68.85; 71.23)	68.26	49.97 (44.91; 55.69)
NPV (%)	67.62 (66.8; 68.63)	66.67	49.98 (44.44; 54.97)	67.67 (66.62; 68.6)	68.82	50.23 (45.29; 55.29)	67.67 (66.71; 68.58)	67.84	49.97 (45.03; 55.56)
AUC (%)	74.88 (73.97; 75.7)	78.05	50 (44.41; 56.22)	74.41 (73.5; 75.22)	77.95	50.15 (44.24; 56.09)	74.82 (73.94; 75.58)	78.38	49.99 (44.11; 56.22)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

raised to 78.9% for tapping data, 75.9% for the multimodal case and 74.6%, 66% and 58.3% for voice, balance and gait data respectively with very high specificities and very low sensitivities.

D. PREDICTIVE FEATURES

Best performance during CV for the main experiment E3 was achieved using the multimodal set of features. Figure 3 shows the scaled average absolute feature weights for RVM and SVM-RFE and the scaled average importance scores for RF, calculated with the out-of-bag (OOB) permuted predictor delta error across 1000 repetitions during the CV. Features with the highest importance scores belong to the tapping task followed by the balance task. Tapping features with the highest importance scores comprised the range of intertap interval (100), maximum value of the intertap interval (99.8) and Teager-Kaiser energy operator of the intertap interval (83.2). Balance features with highest importance scores were the power ratio between high (3.5-15 Hz) and low (0.15-3.5 Hz) frequency for anteroposterior acceleration (31.5) and energy in the medium frequency band for mediolateral acceleration (25.3). Gait and voice tasks had the least contributions in terms of importance scores.

IV. DISCUSSION

Here, we systematically evaluated the ability of four commonly applied DB tasks to differentiate between PD and HC in a self-administered remote setting. Our findings indicate that, depending on the constellation, not accounting for confounds in PD digital biomarker task data may lead to under- but also over-optimistic results.

TABLE 14. Classification performance for the multimodal features.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	73.85 (72.41; 75.47)	76.88	50.04 (45.13; 54.88)	69.39 (67.01; 71.89)	70	50.08 (44.17; 56.67)	69.59 (66.91; 72.43)	73.53	50.04 (43.14; 56.86)
Sensitivity (%)	67.09 (64.57; 69.6)	68.34	38.61 (33.17; 43.97)	66.67 (63.07; 70.12)	65	45.08 (39.17; 51.67)	69.57 (65.69; 73.53)	77.45	53.96 (47.06; 60.78)
Specificity (%)	80.61 (78.59; 82.53)	85.43	61.47 (57.09; 65.79)	72.12 (68.47; 75.52)	75	55.08 (49.17; 61.67)	69.61 (65.69; 73.53)	69.61	46.12 (39.22; 52.94)
PPV (%)	73.57 (71.52; 75.63)	79.07	44.67 (38.37; 50.87)	70.53 (67.71; 73.18)	72.22	50.09 (43.52; 57.41)	69.61 (66.67; 72.86)	71.82	50.04 (43.64; 56.36)
NPV (%)	75.29 (73.92; 76.79)	77.01	55.41 (51.46; 59.31)	68.4 (65.97; 70.97)	68.18	50.07 (44.7; 56.06)	69.6 (66.67; 72.8)	75.53	50.04 (42.55; 57.45)
AUC (%)	82.25 (81.39; 83.15)	85.63	49.96 (44.69; 55.2)	76.01 (74.21; 77.94)	78.81	50.02 (42.89; 57.43)	76.01 (73.79; 78.19)	80.49	50.08 (42.04; 58.23)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	69.24 (66.18; 71.81)	73.04	50.18 (43.63; 56.37)	68.03 (64.95; 70.83)	69.12	50.02 (43.63; 57.35)	69.86 (67.16; 72.79)	70.59	50.01 (43.14; 56.86)
Sensitivity (%)	67.88 (63.73; 71.57)	70.59	47.72 (41.18; 53.92)	65.77 (61.77; 69.61)	63.73	44.62 (38.24; 51.96)	65.98 (62.26; 69.61)	61.76	41.18 (34.31; 48.04)
Specificity (%)	70.6 (66.18; 74.51)	75.49	52.62 (46.08; 58.82)	70.3 (65.69; 74.51)	74.51	55.41 (49.02; 62.75)	73.74 (69.61; 77.94)	79.41	58.83 (51.96; 65.69)
PPV (%)	69.8 (66.5; 72.82)	74.23	50.18 (43.3; 56.7)	68.91 (65.37; 72.21)	71.43	50.02 (42.86; 58.24)	71.56 (68.27; 75.28)	75	50.01 (41.67; 58.33)
NPV (%)	68.75 (65.85; 71.71)	71.96	50.16 (43.93; 56.08)	67.26 (64.39; 70.19)	67.26	50.02 (44.25; 56.64)	68.43 (65.84; 71.27)	67.5	50 (44.17; 55.83)
AUC (%)	74.78 (72.53; 76.97)	80.27	50.04 (42.59; 58.06)	73.49 (71.17; 75.83)	76.68	50.04 (41.42; 58.59)	75.83 (73.72; 78.15)	77.58	50 (42.26; 57.92)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

TABLE 15. Demographics for PD and HC subjects including all data.

	Gait		Balance		Voice		Tapping		Multimodal	
	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC
N	653	2058	655	2092	965	3834	1054	5167	640	1940
Male/ female	436 ± 217	1678 ± 38	438 ± 217	438 ± 385	629 ± 336	3108 ± 726	697 ± 357	4190 ± 977	427 ± 213	1582 ± 358
Age (mean±sd)	60.45 ± 10.72	34.77 ± 14.29	60.44 ± 10.71	34.76 ± 14.23	60.33 ± 11.04	32.77 ± 13.12	59.77 ± 11.42	32.18 ± 12. 53	60.51 ± 10.69	34.84 ± 14.41

A. IDENTIFICATION OF PARKINSON'S DISEASE

Out of the four evaluated machine learning algorithms, similar performance was achieved for all classifiers except LASSO which showed the poorest performance. Whereas some previous studies using the mPower dataset selected different algorithms according to tasks [25], [26], others simply applied a single classifier [27], [29]. No single classifier performed best for all four tasks in our study. This is in line with previous research showing that the selection of the classifier depends mainly on the type and complexity of the data [51], [52]. For instance, RF, RVM and Gaussian SVM are non-linear algorithms, offering more flexibility regarding the

TABLE 16. Classification performance for the gait task.

	Additional experiment: All data			Additional experiment: All data + age + gender		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	50.95 (50.29; 51.63)	51.55	50.01 (49.13; 50.95)	57.2 (55.81; 58.68)	58.25	50.01 (48.54; 51.57)
Sensitivity (%)	2.74 (1.61; 4.13)	3.69	1.35 (0; 2.77)	16.19 (13.3; 19.27)	17.51	5 (2.77; 7.37)
Specificity (%)	99.16 (98.8; 99.6)	99.42	98.68 (98.25; 99.13)	98.21 (97.63; 98.76)	98.98	95.02 (94.32; 95.77)
PPV (%)	50.8 (33.33; 66.67)	66.67	24.34 (0; 50)	74.26 (68.28; 80.66)	84.44	24.12 (13.33; 35.56)
NPV (%)	76.24 (75.99; 76.49)	76.54	75.97 (75.65; 76.32)	78.67 (78.11; 79.26)	79.14	75.97 (75.41; 76.57)
AUC (%)	66.08 (64.67; 67.52)	67.81	49.93 (45.49; 54.46)	86.49 (85.66; 87.32)	88.43	50.01 (45.45; 54.51)

TABLE 17. Classification performance for the balance task.

	Additional experiment: All data			Additional experiment: All data + age + gender		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	58.04(57.02; 59.1)	55.75	49.99 (48.22; 51.53)	69.03 (67.41; 70.53)	65.97	50.07 (47.6; 52.41)
Sensitivity (%)	19.12 (17.16; 21.28)	14.22	5.46 (2.75; 7.8)	41.9 (38.9; 45.08)	36.24	12.02 (8.26; 15.6)
Specificity (%)	96.96 (96.34; 97.56)	97.27	94.53 (93.69; 95.27)	96.16 (95.56; 96.77)	95.7	88.12 (86.94; 89.24)
PPV (%)	66.35 (61.03; 71.31)	62	23.78 (12; 34)	77.36 (74.48; 80.26)	72.48	24.05 (16.51; 31.19)
NPV (%)	79.28 (78.88; 79.71)	78.38	76.17 (75.49; 76.76)	84.09 (83.35; 84.8)	82.75	76.21 (75.19; 77.17)
AUC (%)	73.42 (72.51; 74.34)	72.05	49.98 (45.45; 54.5)	89.46 (88.93; 90.02)	89.57	49.99 (45.79; 54.5)

type of data. On the contrary, LASSO is a linear classifier and thus, its performance depends on whether the data is linearly separable. Whereas the generalizability of this observation is limited by the use of only one linear classifier, it may point to a better usability of non-linear approaches for classification of digital assessments.

For discrimination of PD and HC, combination of all tasks reached a BA of 74%, followed by tapping that achieved 67%, outperforming other tasks which were close to chance level. These results are in line with previous literature using the mPower dataset, where tapping reached the highest accuracies and gait and voice were closer to chance level [29]. Several studies reported higher accuracies for this type of data [24], [27]. Yet, these studies followed certain ''optimistic'' approaches as discussed below.

B. POTENTIAL CONFOUNDERS

Exclusion of comorbidities resulted in increased accuracies by a few percent, suggesting that other diseases may add more variability to the signal. Prediction performances considerably decreased for all tasks after matching for age and sex indicating the importance of controlling for such confounds in DB data. When including all data without trimming age range, accuracies greatly increase. Nonetheless, specificity

TABLE 18. Classification performance for the voice task.

	Additional experiment: All data			Additional experiment: All data + age + gender		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	52.91 (52.13; 53.68)	52.57	50.03 (49.26; 51.01)	73.12 (71.9; 74.35)	74.6	50 (48.09; 52.19)
Sensitivity (%)	7.59 (6.06; 9.16)	6.23	2.17 (0.94; 3.74)	50.68 (48.29; 53.11)	53.89	14.58 (11.53; 18.07)
Specificity (%)	98.24 (97.93; 98.55)	98.9	97.89 (97.57; 98.28)	95.55 (95.11; 96.01)	95.31	85.43 (84.66; 86.31)
PPV (%)	52.05 (45.24; 58.7)	58.82	20.49 (8.82; 35.29)	74.18 (72.28; 76.2)	74.25	20.08 (15.88; 24.89)
NPV (%)	80.84 (80.58; 81.1)	80.77	79.93 (79.68; 80.26)	88.49 (88; 89)	89.17	79.93 (79.21; 80.75)
AUC (%)	73.23 (72.43; 74.01)	74.91	50.09 (46.63; 53.52)	91.39 (91.05; 91.72)	91.16	49.99 (46.49; 53.58)

TABLE 19. Classification performance for the tapping task.

	Experiment 7			Experiment 8		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	74.42 (73.79; 74.99)	71.08	50.03 (48.27; 51.87)	80.81 (79.99; 81.58)	78.91	50 (48.22; 51.82)
Sensitivity (%)	52.65 (51.49; 53.77)	45.87	10.9 (7.98; 13.96)	65.68 (64.01; 67.28)	61.25	13.22 (10.26; 16.24)
Specificity (%)	96.18 (95.91; 96.43)	96.28	89.16 (88.56; 89.78)	95.94 (95.68; 96.24)	96.57	86.78 (86.18; 87.4)
PPV (%)	73.75 (72.41; 75.15)	71.56	17.01 (12.44; 21.78)	76.76 (75.53; 78.17)	78.47	16.94 (13.14; 20.8)
NPV (%)	90.87 (90.66; 91.07)	89.72	83.08 (82.52; 83.66)	93.2 (92.9; 93.49)	92.44	83.07 (82.49; 83.66)
AUC (%)	89.39 (89.08; 89.71)	88.64	50.04 (46.6; 53.56)	94.51 (94.29; 94.72)	94.78	50.12 (46.59; 53.37)

values are exceedingly high whereas sensitivity values are vastly low. This indicates a greater prediction ability for the HC group, which is considerably larger than the PD group for subjects under 35 years old. Including age and sex as part of the features resulted in further accuracy increases, yet with very low sensitivities. Since the dataset is strongly slanted toward young HC, the model is most likely distinguishing HC based on age and gender in this case. Such effects may also explain the high accuracies in some of the previous studies using mPower dataset, where no proper matching for these confounds was performed, age and/or sex were used as features despite a large imbalance across groups or non-balanced accuracies were reported [24], [26], [27], [34]. In example, in the overall mPower dataset HC outnumber PD by a factor of five and age and sex alone provide a high discrimination accuracy between PD and HC with PD being on average 28 years older and more often female (34% of PD vs 19% of HC). Our findings are also in line with previous studies demonstrating a similarly strong decrease in accuracies when accounting for respective confounds. Neto *et al.* [53] studied the effect of confounders on gait data. They reached very high accuracy when not accounting for confounders, compared with a very modest accuracy when using unconfounded measures. Schwab and Karlent [25] performed analysis with all

TABLE 20. Classification performance for the multimodal features.

	Experiment 7			Experiment 8		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	72.67 (71.82; 73.48)	73.26	50.07 (47.2; 52.97)	75.34 (74.44; 76.19)	75.88	50.06 (47.32; 53.06)
Sensitivity (%)	49.44 (47.78; 50.94)	49.77	14.89 (10.8; 19.25)	54.66 (53.05; 56.32)	55.63	16.79 (12.68; 21.13)
Specificity (%)	95.91 (95.52; 96.33)	96.75	85.25 (83.75; 86.69)	96.03 (95.67; 96.41)	96.13	83.32 (81.89; 84.83)
PPV (%)	79.95 (78.26; 81.67)	83.46	24.97 (17.83; 32.28)	81.96 (80.51; 83.51)	82.58	24.93 (18.82; 31.69)
NPV (%)	85.18 (84.77; 85.57)	85.38	75.23 (73.98; 76.5)	86.52 (86.08; 86.95)	86.79	75.23 (74; 76.57)
AUC (%)	88.99 (88.46; 89.48)	89.22	50.15 (45.55; 54.75)	92.39 (91.97; 92.81)	92.15	50.03 (45.74; 54.63)

the tasks from the mPower dataset with and without including age and sex, the latter resulting in a similarly low accuracy as in our study.

For all classification experiments, we used only one recording per subject to prevent the classifier from detecting the idiosyncrasies of each subject rather than specific PD related symptoms [29]–[31]. Single measures are likely to contain more noise due to higher variation in task administration as well as in individual performance in a poorly-controlled setting [54]. Using multiple time points may therefore further increase the discrimination between PD and HC as demonstrated in several previous studies [29]–[31]. Yet, our results in this respect highlight the need of further understanding and better control of the individual parameters which impact the task performance during a single administration.

C. PREDICTORS OF PARKINSON'S DISEASE

Features with largest weights in the multimodal discrimination between PD and HC were derived from the tapping task. These features mostly related to the inter-tapping interval (time), presumably reflecting bradykinesia-like symptoms. These results are in line with previous studies, where tapping features related to speed and accuracy had the strongest correlation with clinical scores [55], [56]. Balance task features related to tremor measures had larger weights than postural ones. In addition, features from the frequency domain had greater weights than spatiotemporal features. Spatiotemporal features have been extensively studied and applied, due to their ease of computation and interpretability [57]. However, these features offer information limited primarily to leg movement, whilst frequency features add information regarding asymmetry and variability. Furthermore, balance features with higher weights belonged to the mediolateral and anteroposterior signals, related to stability. Even though gait had limited contribution to the classification accuracy, acceleration features had the highest weights from this task. This observation is in line with previous findings where acceleration proved to better capture PD-related gait changes [58]. In line with some previous studies, features with the highest weights from the voice task were all based on Mel Frequency Cepstral Coefficients which can detect subtle changes in speech articulation that are common in PD [59], [60].

D. LIMITATIONS AND FURTHER RESEARCH

Whereas sensors-integrated in smartphones open new opportunities for at-home continuous, reliable, non-invasive and low-cost monitoring of PD, our finding highlight the need for further development, optimization and standardization of specific measures for such applications.

The interpretation of our findings is limited by several aspects, including the lack of standardization, poor control of environmental and medication effects during performance of the tasks and intentionally or unintentionally incorrect information provided by the participants. In addition, removal of comorbidities and matching for age and sex led to exclusion of about 50% of data, which may affect the training of classifiers [53].

Further use of smartphones in the detection of Parkinson's disease symptoms include detection of hypomimia from face expressions, socializing and lifestyle behavior and typing patterns among others [61], [62].

APPENDIX A
SUPPLEMENTARY METHODS

A. DATA CLEANING

MPower dataset offers demographic, PDQ8 and MDS-UPDRS surveys and task-based data. The demographics table contains data for 6805 subjects. In order to establish a diagnosis, participants had to select ''true'' or ''false'' to the following question ''Have you been diagnosed by a medical professional with Parkinson Disease?''. According to this answer, they are classified as Parkinson's Disease (PD) or Healthy Control (HC). Some subjects left this question unanswered and thus they were discarded from further analysis. Those subjects classified as PD which did not completed the PDQ8 and MDS-UPDRS questionnaire were also excluded. Subjects with no information on age, sex or any task data were also removed, resulting in 6614 subjects. Those empty, null or corrupted files for each task were deleted, resulting in 2807 subjects with gait and balance data, 4925 with voice data and 6366 with tapping data. Since a large number of subjects are HC under 35 years old, our analysis focused on a subset of subjects within the age range of 35 to 75 years old, leading to 1435 subjects with gait and balance data, 2186 subjects with voice data and 2644 subjects with tapping data. Finally, all subjects with inconsistencies for each of the tasks were discarded (i.e., subjects that reported not to have been diagnosed with Parkinson's disease but filled in PD medication questions, year of diagnosis of PD, surgery or deep brain stimulation). This last elimination resulted in 1416 subjects with gait and balance data, 2153 subjects with voice data and 2600 subjects with tapping data.

B. SIGNALS LENGTH

Gait task consists of walking 20 steps in a straight line. In order to analyse the same signal length for each subject, we examined how many subjects had gait data for different time durations. We observed that after 10 seconds, participation was dropping heavily. Therefore, we selected a time length of 10 seconds and discarded those participants with shorter signals. Following the same reasoning, we chose voice signals of 7 seconds, trimming the first second and last two seconds, and tapping signals of 20 seconds. Similarly, balance task consists of standing still for 30 seconds although just 20 seconds were selected. Nonetheless, whereas gait, voice and tapping are independent tasks, and therefore they are started by the user, balance task starts straight after the gait task. This is, as soon as the gait task ends, the app plays out loud ''turn around and stand still for 30 seconds''. As a result, most of the balance recordings include initial slots of noise, which most likely coincide with the time that subjects listen to the instructions, react, turn around and start the task. Therefore, we trimmed the first 5 seconds of the signal, resulting in balance signals of 15 seconds for all subjects. Final number of subjects consisted of 1397 subjects with gait data, 1415 subjects with balance data, 2150 subjects with voice data and 2600 subjects with tapping data.

C. PRE-PROCESSING AND SIGNAL EXTRACTION

Gait and balance data consists on vertical (V), anteroposterior (AP) and mediolateral (ML) acceleration signals. For these 3 gait acceleration signals, we applied a Butterworth low pass filter with cut-off frequency at 20 Hz followed by a 3° order high pass filter at 0.3 Hz. According to Pittman *et al.* [24], around 30% of devices were not held in the correct position. Therefore, the greatest gravitational displacement is not always along the vertical axis. Then, we standardized these three signals and calculated an additional average acceleration signal. Based on the standardized acceleration signal, we extracted the step series. We calculated position signals along the three axes by double integrating the acceleration signals and the average position. Then, we extracted velocity and acceleration along the path by derivation [37].

Balance acceleration signals were filtered with a low pass Butterworth filter at 20 Hz. Since tremor in PD usually falls in the 4-7Hz frequency band [38], [39], the interval 0-3.5 Hz is considered for tremor-free or postural acceleration measures. Hence, we applied a Butterworth filter at 3.5 Hz to extract postural acceleration measures. We also calculated the average of the tremor acceleration in the 3 axes and the average of the postural acceleration in the 3 axes.

Voice signals were recorded at a sample frequency of 44.1 KHz. We downsampled the signal to 25KHz, applied a second order Butterworth filter with cut-off frequency at 400 Hz followed by a pre-emphasis FIR filter for noise reduction and correct for distortions. We extracted the fundamental frequency (f0) series, which was verified with the Praat software.

Tapping recordings consists of the {x,y} screen pixel coordinates and timestamp for each tap on the screen. Signals derived out of these recordings were the inter-tapping interval (time) and the {x,y} inter-tap distance series.

D. FEATURE EXTRACTION

1) GAIT
We extracted 11 signals from the original accelerometer recordings during gait tasks. These are V, AP and ML acceleration, the step series, the average of the acceleration in the three axes, the V, AP and ML position, the average position in the three axes, the velocity and the acceleration along the path. Table 4 collects a list of features extracted for these signals along with their acronyms.

2) BALANCE
Balance signals consist in the V, AP and ML tremor acceleration (4-7 Hz), the average of these 3 signals, the V, AP and ML postural acceleration (0-3.5 Hz) and the average of these 3 signals. We extracted displacement-related postural features from ML, AP and average of both distance signals, following the formulation in Martinez-Mendez *et al.* [36] (Table 5).

3) VOICE
Most of voice features were extracted following the formulation in Tsanas *et al.* [45]. Tsanas *et al.* state that the period (T) signal provides different information than f0. Therefore, we additionally extracted the T series. Further signals include glottis quotient and 14 Mel Frequency Cepstral Coefficients (MFCCs),

including the 0th coefficient and the log-energy of the signal, along with their associated delta and delta-delta coefficients as applied in the Voicebox Matlab ToolBox [63] (Table 6).

4) TAPPING
We considered a set of features computed from the inter-tapping interval (time) and the {x,y} inter-tap distance signals, according to Bot *et al.* [46] (Table 7).

E. COMORBIDITIES

Comorbidities selected for removal in the experiments E3-E6 include ''Alzheimer Disease or Alzheimer dementia'', ''Dementia'', ''Schizophrenia or Bipolar Disorder'', ''Alcoholism'', ''Multiple Sclerosis'', ''Leukemia or Lymphoma'', ''Acute Myocardial Infarction/Heart Attack'', ''Stroke/Transient Ischemic Attack'', ''Breast Cancer'', ''Colorectal Cancer'', ''Prostate Cancer'', ''Lung Cancer'', ''Endometrial/Uterine Cancer'', ''Any other kind of cancer OR tumor'', ''Heart Failure/Congestive Heart Failure'', ''Ischemic Heart Disease''. These comorbidities were removed since they may lead to brain damage or to undertake chemotherapy or other therapy, which might induce brain changes.

F. MEDICATION STATUS

Table 8 shows the number of subjects that performed the task just before taking their medication, after taking their medication, at another random time, number of those who were not taking any medication and number of those who did not give any information about their medication status.

G. SELECTION OF THE BEST CLASSIFIER DURING THE MAIN EXPERIMENT (NO COMORBIDITIES; MATCHED)

Table 9 shows the classification performance for the four classifiers under consideration for each task.

APPENDIX B SUPPLEMENTARY RESULTS

Table 10-Table 14 summarize the results for each task (gait, balance, voice, tapping) and the combination of all the tasks, for the experiment 1 (all data), experiment 2 (matched data), experiment 3 (no comorbidities and matched data), experiment 4 (no comorbidities, matched, controlled for age), experiment 5 (no comorbidities, matched, controlled for sex) and experiment 6 (no comorbidities, matched, controlled for age and sex).

A. ADDITIONAL EXPERIMENTS

Our results may differ to those in the current literature using the mPower dataset since we follow different approaches. To explain these discrepancies and compare with the literature, we included two additional experiments including all data without trimming for age range and all data including both age and sex as features in the analyses (Table 15). Classification performances for both additional experiments for each tasks are summarized in Table 16-Table 20.

ACKNOWLEDGMENT

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

María Goñi performed the overall analyses and wrote manuscript. Juergen Dukart designed the overall study and contributed to writing the manuscript. Kaustubh R. Patil, Simon B. Eickhoff, and Mehran Sahandi Far provided input on the analyses. All authors contributed to interpretation of results, reviewed, and commented on the manuscript.

AUTHOR DECLARATION

The access to the Mpower data was granted after registration in the Synapse system, signing an oath, submitting an Intended Data Use Statement and accepting data-specific Conditions. All appropriate institutional forms have been archived.

REFERENCES

[1] C. H. Adler, T. G. Beach, J. G. Hentz, H. A. Shill, J. N. Caviness, E. Driver-Dunckley, M. N. Sabbagh, L. I. Sue, S. A. Jacobson, C. M. Belden, and B. N. Dugger, ''Low clinical diagnostic accuracy of early vs advanced Parkinson disease: Clinicopathologic study,'' *Neurology*, vol. 83, no. 5, pp. 406–412, Jul. 2014.

[2] J.-W. Kim, Y. Kwon, Y.-M. Kim, H.-Y. Chung, G.-M. Eom, J.-H. Jun, J.-W. Lee, S.-B. Koh, B. K. Park, and D.-K. Kwon, ''Analysis of lower limb bradykinesia in Parkinson's disease patients,'' *Geriatrics Gerontology Int.*, vol. 12, no. 2, pp. 257–264, Apr. 2012.

[3] J.-F. Daneault, S. I. Lee, F. N. Golabchi, S. Patel, L. C. Shih, S. Paganoni, and P. Bonato, ''Estimating bradykinesia in Parkinson's disease with a minimum number of wearable sensors,'' in *Proc. IEEE/ACM Int. Conf. Connected Health: Appl., Syst. Eng. Technol. (CHASE)*, Jul. 2017, pp. 264–265.

[4] H. Zach, A. M. Janssen, A. H. Snijders, A. Delval, M. U. Ferraye, E. Auff, V. Weerdesteyn, B. R. Bloem, and J. Nonnekes, ''Identifying freezing of gait in parkinson's disease during freezing provoking tasks using waist-mounted accelerometry,'' *Parkinsonism Rel. Disorders*, vol. 21, no. 11, pp. 1362–1366, Nov. 2015.

[5] A. Suppa, A. Kita, G. Leodori, A. Zampogna, E. Nicolini, P. Lorenzi, R. Rao, and F. Irrera, ''L-DOPA and freezing of gait in Parkinson's disease: Objective assessment through a wearable wireless system,'' *Frontiers Neurol.*, vol. 8, p. 406, Aug. 2017.

[6] N.-H. Ko, C. M. Laine, B. E. Fisher, and F. J. Valero-Cuevas, ''Force variability during dexterous manipulation in individuals with mild to moderate Parkinson's disease,'' *Frontiers Aging Neurosci.*, vol. 7, p. 151, Aug. 2015.

[7] R. P. Hubble, G. A. Naughton, P. A. Silburn, and M. H. Cole, ''Wearable sensor use for assessing standing balance and walking stability in people with Parkinson's disease: A systematic review,'' *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0123705.

[8] H. Dubey, J. C. Goldberg, M. Abtahi, L. Mahler, and K. Mankodiya, ''EchoWear: Smartwatch technology for voice and speech treatments of patients with Parkinson's disease,'' 2016, *arXiv:1612.07608*.

[9] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, ''Fully automated assessment of the severity of Parkinson's disease from speech,'' *Comput. Speech Lang.*, vol. 29, no. 1, pp. 172–185, Jan. 2015.

[10] A. J. Espay, P. Bonato, F. B. Nahab, W. Maetzler, J. M. Dean, J. Klucken, B. M. Eskofier, A. Merola, F. Horak, A. E. Lang, and R. Reilmann, ''Technology in parkinson's disease: Challenges and opportunities,'' *Movement Disorders*, vol. 31, no. 9, pp. 1272–1282, Apr. 2016.

[11] E. Rovini, C. Maremmani, and F. Cavallo, ''How wearable sensors can support Parkinson's disease diagnosis and treatment: A systematic review,'' *Frontiers Neurosci.*, vol. 11, p. 555, Oct. 2017.

[12] M. Linares-del Rey, L. Vela-Desojo, and R. Cano-de la Cuerda, ''Mobile phone applications in Parkinson's disease: A systematic review,'' *Neurología*, vol. 34, no. 1, pp. 38–54, Jan. 2019.

[13] W. Maetzler, J. Domingos, K. Srulijes, J. J. Ferreira, and B. R. Bloem, ''Quantitative wearable sensors for objective assessment of Parkinson's disease,'' *Movement Disorders*, vol. 28, no. 12, pp. 1628–1637, Oct. 2013.

[14] S. Arora, V. Venkataraman, S. Donohue, K. M. Biglan, E. R. Dorsey, and M. A. Little, ''High accuracy discrimination of Parkinson's disease participants from healthy controls using smartphones,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3641–3644.

[15] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. M. Biglan, E. R. Dorsey, and M. A. Little, ''Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study,'' *Parkinsonism Rel. Disorders*, vol. 21, no. 6, pp. 650–653, Jun. 2015.

[16] A. Benba, A. Jilbab, and A. Hammouch, ''Detecting patients with Parkinson's disease using Mel frequency cepstral coefficients and support vector machines,'' *Int. J. Electr. Eng. Inform.*, vol. 7, no. 2, pp. 297–307, Jun. 2015.

[17] N. Kostikis, D. Hristu-Varsakelis, M. Arnaoutoglou, and C. Kotsavasiloglou, ''A smartphone-based tool for assessing parkinsonian hand tremor,'' *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1835–1842, Nov. 2015.

[18] M. Suzuki, H. Mitoma, and M. Yoneyama, ''Quantitative analysis of motor status in Parkinson's disease using wearable devices: From methodological considerations to problems in clinical applications,'' *Parkinsons Dis*, vol. 2017, May 2017, Art. no. 6139716.

[19] K. Szewczyk-Krolikowski, P. Tomlinson, K. Nithi, R. Wade-Martins, K. Talbot, Y. Ben-Shlomo, and M. T. M. Hu, ''The influence of age and gender on motor and non-motor features of early Parkinson's disease: Initial findings from the Oxford Parkinson disease center (OPDC) discovery cohort,'' *Parkinsonism Rel. Disorders*, vol. 20, no. 1, pp. 99–105, Jan. 2014.

[20] M. Picillo, A. Nicoletti, V. Fetoni, B. Garavaglia, P. Barone, and M. T. Pellecchia, ''The relevance of gender in Parkinson's disease: A review,'' *J. Neurol.*, vol. 264, no. 8, pp. 1583–1607, Aug. 2017.

[21] S. Nazem, A. D. Siderowf, J. E. Duda, T. T. Have, A. Colcher, S. S. Horn, P. J. Moberg, J. R. Wilkinson, H. I. Hurtig, M. B. Stern, and D. Weintraub, ''Montreal cognitive assessment performance in patients with Parkinson's disease with 'normal' global cognition according to mini-mental state examination score,'' *J. Amer. Geriatrics Soc.*, vol. 57, no. 2, pp. 304–308, Feb. 2009.

[22] M. M. Wickremaratchi, M. D. W. Knipe, B. S. D. Sastry, E. Morgan, A. Jones, R. Salmon, R. Weiser, M. Moran, D. Davies, L. Ebenezer, S. Raha, N. P. Robertson, C. C. Butler, Y. Ben-Shlomo, and H. R. Morris, ''The motor phenotype of Parkinson's disease in relation to age at onset,'' *Movement Disorders*, vol. 26, no. 3, pp. 457–463, Feb. 2011.

[23] L. M. Shulman, R. L. Taback, J. Bean, and W. J. Weiner, ''Comorbidity of the nonmotor symptoms of Parkinson's disease,'' *Movement Disorders, Official J. Movement Disorder Soc.*, vol. 16, no. 3, pp. 507–510, 2001.

[24] B. Pittman, R. H. Ghomi, and D. Si, ''Parkinson's disease classification of mPower walking activity participants,'' in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 4253–4256.

[25] P. Schwab and W. Karlen, ''PhoneMD: Learning to diagnose Parkinson's disease from smartphone data,'' 2018, *arXiv:1810.01485*.

[26] J. Prince, F. Andreotti, and M. De Vos, ''Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data,'' *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1402–1411, May 2019.

[27] S. Mehrang, M. Jauhiainen, J. Pietila, J. Puustinen, J. Ruokolainen, and H. Nieminen, ''Identification of Parkinson's disease utilizing a single self-recorded 20-step walking test acquired by Smartphone's inertial measurement unit,'' in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 2913–2916.

[28] M. Memedi, A. Sadikov, V. Groznik, J. Žabkar, M. Možina, F. Bergquist, A. Johansson, D. Haubenberger, and D. Nyholm, ''Automatic spiral analysis for objective assessment of motor symptoms in Parkinson's disease,'' *Sensors*, vol. 15, no. 9, pp. 23727–23744, Sep. 2015.

[29] E. C. Neto, T. M Perumal, A. Pratap, B. M Bot, L. Mangravite, and L. Omberg, ''On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: The mPower case study,'' 2017, *arXiv:1706.09574*.

[30] E. C. Neto, A. Pratap, T. M. Perumal, M. Tummalacherla, P. Snyder, B. M. Bot, A. D. Trister, S. H. Friend, L. Mangravite, and L. Omberg, ''Detecting the impact of subject characteristics on machine learning-based diagnostic applications,'' *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–6, Oct. 2019.

[31] E. C. Neto, A. Pratap, T. M. Perumal, M. Tummalacherla, B. M. Bot, A. D Trister, S. H Friend, L. Mangravite, and L. Omberg, ''Learning disease vs participant signatures: A permutation test approach to detect identity confounding in machine learning diagnostic applications,'' 2017, *arXiv:1712.03120*.

[32] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, and A. D. Trister, ''The mPower study, Parkinson disease mobile data collected using ResearchKit,'' *Sci. Data*, vol. 3, no. 1, pp. 1–9, Mar. 2016.

[33] A. Zhan, S. Mohan, C. Tarolli, R. B. Schneider, J. L. Adams, S. Sharma, M. J. Elson, K. L. Spear, A. M. Glidden, M. A. Little, and A. Terzis, ''Using smartphones and machine learning to quantify Parkinson disease severity: The mobile Parkinson disease score,'' *JAMA Neurol.*, vol. 75, no. 7, pp. 876–880, Jul. 2018.

[34] M. Giuliano, A. García-López, S. Pérez, F. D. Pérez, O. Spositto, and J. Bossero, ''Selection of voice parameters for Parkinson's disease prediction from collected mobile data,'' in *Proc. 22nd Symp. Image Signal Process. Artif. Vis. (STSIVA)*, Apr. 2019, pp. 1–3.

[35] J. Prince and M. de Vos, ''A deep learning framework for the remote detection of Parkinson'S disease using smart-phone sensor data,'' in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 3144–3147.

[36] R. Martinez-Mendez, M. Sekine, and T. Tamura, ''Postural sway parameters using a triaxial accelerometer: Comparing elderly and young healthy adults,'' *Comput. Methods Biomech. Biomed. Eng.*, vol. 15, no. 9, pp. 899–910, Sep. 2012.

[37] K. Seifert and O. Camacho, ''Implementing positioning algorithms using accelerometers,'' *Freescale Semicond.*, vol. 1, p. 13, Feb. 2007.

[38] K. E. Lyons, R. Pahwa, and R. Pahwa, *Handbook of Essential Tremor and Other Tremor Disorders*. Boca Raton, FL, USA: CRC Press, 2005.

[39] L. Palmerini, L. Rocchi, S. Mellone, F. Valzania, and L. Chiari, ''Feature selection for accelerometer-based posture analysis in Parkinson's disease,'' *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 3, pp. 481–490, May 2011.

[40] A. Zhan, M. A. Little, D. A. Harris, S. O. Abiola, E. R. Dorsey, S. Saria, and A. Terzis, ''High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection,'' 2016, *arXiv:1601.00960*.

[41] A. Weiss, S. Sharifi, M. Plotnik, J. P. P. van Vugt, N. Giladi, and J. M. Hausdorff, ''Toward automated, at-home assessment of mobility among patients with Parkinson disease, using a body-worn accelerometer,'' *Neurorehabilitation Neural Repair*, vol. 25, no. 9, pp. 810–818, Nov. 2011.

[42] R. San-Segundo, R. Torres-Sánchez, J. Hodgins, and F. De la Torre, ''Increasing robustness in the detection of freezing of gait in Parkinson's disease,'' *Electronics*, vol. 8, no. 2, p. 119, Jan. 2019.

[43] M. Bachlin, M. Plotnik, D. Roggen, I. Maidan, J. M. Hausdorff, N. Giladi, and G. Troster, ''Wearable assistant for Parkinson's disease patients with the freezing of gait symptom,'' *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 436–446, Mar. 2010.

[44] T. E. Prieto, J. B. Myklebust, R. G. Hoffmann, E. G. Lovett, and B. M. Myklebust, ''Measures of postural steadiness: Differences between healthy young and elderly adults,'' *IEEE Trans. Biomed. Eng.*, vol. 43, no. 9, pp. 956–966, Sep. 1996.

[45] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, ''Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity,'' *J. Roy. Soc. Interface*, vol. 8, no. 59, pp. 842–855, Jun. 2011.

[46] B. M. Bot. *mPower: Public Researcher Portal*. Accessed: Jun. 25, 2020. [Online]. Available: https://www.synapse.org/#!Synapse:syn4993293/files/

[47] R. Tibshirani, ''Regression shrinkage and selection via the lasso,'' *J. Roy. Statist. Soc., B*, vol. 58, no. 1, pp. 267–288, 1996.

[48] L. Breiman, ''Random forests,'' *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[49] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, ''Gene selection for cancer classification using support vector machines,'' *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, Jan. 2002.

[50] M. E. Tipping, ''Sparse Bayesian learning and the relevance vector machine,'' *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jun. 2001.

[51] S. Bind, A. K. Tiwari, and A. K. Sahani, "A survey of machine learning based approaches for Parkinson disease prediction," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 1648–1655, 2015.

[52] P. B. Brazdil, C. Soares, and J. P. da Costa, "Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results," *Mach. Learn.*, vol. 50, no. 3, pp. 251–277, Mar. 2003.

[53] E. C. Neto, A. Pratap, T. M Perumal, M. Tummalacherla, B. M Bot, L. Mangravite, and L. Omberg, "Using permutations to assess confounding in machine learning applications for digital health," 2018, *arXiv:1811.11920*.

[54] M. S. Far, S. B. Eickhoff, M. Goñi, and J. Dukart, "Exploring test retest reliability and longitudinal stability of digital biomarkers for Parkinson's disease in the m-power dataset: Cohort study," *J. Med. Internet Res.*, vol. 23, no. 9, p. e26608, Sep. 2021.

[55] M. Memedi, T. Khan, P. Grenholm, D. Nyholm, and J. Westin, "Automatic and objective assessment of alternating tapping performance in Parkinson's disease," *Sensors*, vol. 13, no. 12, pp. 16965–16984, Dec. 2013.

[56] C. Y. Lee, S. J. Kang, S.-K. Hong, H.-I. Ma, U. Lee, and Y. J. Kim, "A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease," *PLoS ONE*, vol. 11, no. 7, Jul. 2016, Art. no. e0158852.

[57] F. Wahid, R. K. Begg, C. J. Hass, S. Halgamuge, and D. C. Ackland, "Classification of parkinson's disease gait using spatial-temporal gait features," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 6, pp. 1794–1802, Nov. 2015.

[58] E. Sejdic, K. A. Lowry, J. Bellanca, M. S. Redfern, and J. S. Brach, "A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 3, pp. 603–612, May 2014.

[59] T. Khan. Running-Speech MFCC are Better Markers of Parkinsonian Speech Deficits Than Vowel Phonation and Diadochokinetic. DiVA. Accessed: Dec. 2020. [Online]. Available: http://mdh.diva-portal.org/smash/record.jsf?pid=diva2%3A705196&dswid=6494/

[60] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[61] Y. Seliverstov, D. Diagovchenko, M. Kravchenko, M. Babin, E. Fedotova, and M. Belyaev, "Hypomimia detection with a smartphone camera as a possible self-screening tool for Parkinson disease," *J. Neurol.*, vol. 90, no. 15, p. 3.047, 2018.

[62] R. B. Schneider, L. Omberg, E. A. Macklin, M. Daeschler, L. Bataille, S. Anthwal, T. L. Myers, E. Baloga, S. Duquette, P. Snyder, and K. Amodeo, "Design of a virtual longitudinal observational study in Parkinson's disease (AT-HOME PD)," *Ann. Clin. Transl. Neurol.*, vol. 8, no. 2, pp. 308–320, Feb. 2020.

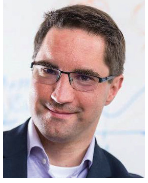
[63] M. Brookes. *VOICEBOX: Speech Processing Toolbox for MATLAB*. Accessed: Dec. 2020. [Online]. Available: http://www.ee.ic.ac.U.K./hp/staff/dmb/voicebox/voicebox.html/

[64] A. Mirelman, T. Heman, K. Yasinovsky, A. Thaler, T. Gurevich, K. Marder, S. Bressman, A. Bar-Shira, A. Orr-Urtreger, N. Giladi, and J. M. Hausdorff, "Fall risk and gait in Parkinson's disease: The role of the LRRK2 G2019S mutation," *Movement Disorders*, vol. 28, no. 12, pp. 1683–1690, Oct. 2013.

[65] B. M. Bot. *Sage-Bionetworks: MPower-Sdata*. Accessed: Jul. 2020. [Online]. Available: https://github.com/Sage-Bionetworks/mPower-sdata/

MARÍA GOÑI received the B.S. degree in technical industrial engineering from the University of Cantabria, Spain, in 2008, the M.S. degree in biomedical engineering from the University of Pais Vasco, Spain, in 2014, the Ph.D. degree in neurosciences from the University of Aberdeen, U.K., in 2019, and the B.S. degree in electronics engineering from the University of Alcalá, Spain, in 2021.

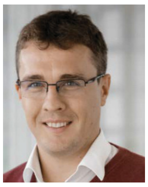
Since 2019, she has been a Postdoctoral Researcher with the Institute of Neuroscience and Medicine, Research Centre Jülich, Germany. Her research interests include the identification of prognostic biomarkers in neuropsychiatric diseases by integrating different neuroimaging modalities and sensor-based data and the application of machine learning and other analytical techniques.



SIMON B. EICKHOFF is currently a Full Professor and the Chair of the Institute for Systems Neuroscience, Heinrich Heine University, Düsseldorf, and the Director of the Institute of Neuroscience and Medicine (INM-7, Brain and Behavior), Forschungszentrum Jülich. He is a Visiting Professor at the Institute of Automation, Chinese Academy of Sciences. He is working at the interface between neuroanatomy, data-science, and brain medicine. He aims to obtain a more detailed characterization of the organization of the human brain and its inter-individual variability in order to better understand its changes in advanced age and neurological and psychiatric disorders. This goal is pursued by the development and application of novel analysis tools and approaches for large-scale, multi-modal analysis of brain structure, function and connectivity and by machine-learning for single subject prediction of cognitive and socio-affective traits, and ultimately precision medicine.

MEHRAN SAHANDI FAR received the master's degree in computer science from Eastern Mediterranean University. He is currently pursuing the Ph.D. degree with the Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, and a Researcher with the Research Center Jülich's Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour). His current research interests include digital biomarker development and remote monitoring using smart devices.

KAUSTUBH R. PATIL (Member, IEEE) received the B.E. degree in electronics engineering from Shivaji University, India, in 2003, the M.Sc. degree in artificial intelligence and intelligent systems from the University of Porto, Portugal, in 2007, and the Ph.D. degree from the Max Planck Institute of Computer Science, Germany, in 2013. From 2013 to 2016, he was a Postdoctoral Fellow at UCL and MIT. He joined FZJ, in 2017, where he is currently leading the Applied Machine Learning Group. He works on the application of machine learning techniques to better understand biological systems. He is an Associate Editor of IEEE Access.



JUERGEN DUKART received the Diploma degree in psychology from Ruhr University Bochum, in 2008, and the Ph.D. degree in neuroscience from the Max Planck Institute for Human Cognitive and Brain Sciences, in 2011. He was a Postdoctoral Researcher at the University of Lausanne, Switzerland. Then, he moved to the Pharmaceutical Company F. Hoffmann-La Roche, where he worked for five years in different functions, such as the Head of Clinical Imaging and a Biomarker Experimental Medicine Leader being responsible for imaging and overall biomarker strategy in various phase I to phase III clinical trials. Since 2019, he has been leading the Group Biomarker Development, Institute of Neuroscience and Medicine (INM-7), Research Centre Jülich, Germany. He is the author of more than 60 articles. His main research interests include the development of technology-based biomarkers for early detection, follow-up, stratification, and monitoring of treatment effects in neurological and psychiatric diseases. He is an Associate Editor of the journal *Frontiers in Human Neuroscience*.

General Discussion

DBs collected using smart devices are a promising tool for accurate, continuous, and unobstructed monitoring of PD symptoms in patients' daily life. Nonetheless, for DBs to be successfully implemented in clinical settings, it is crucial to establish an open-source infrastructure, standardised evaluation pipelines, and the development of reliable DBs. In a series of manuscripts, I addressed some of the known major challenges in this context, and in the following section, I discuss the key outcomes.

The primary objective of the work presented in the first manuscript was to develop an open-source, modular platform for digital phenotyping with a focus on privacy and security (68). The resulting solution, "JTrack", was developed based on these criteria to facilitate unobstructed and remote sensor-based psychiatric and neurological disorders monitoring. At the same time, it addresses the major shortcomings of existing platforms. "JTrack" complies with GDPR security and privacy regulations, making it trustworthy and available through official application stores (Google Play and Apple Store). This is in contrast to some evaluated platforms that lack this level of trust and security as they rely on online distribution methods.

Additionally, "JTrack" reduces the impact of technical failures and missing data through optimal battery and memory use. Lastly, while the motor manifestations of neurodegenerative diseases, particularly PD, are the most studied, the non-motor manifestations are barely studied using remote monitoring methods, owing to the difficulty of collecting and integrating relative modules in current platforms. In this regard, the modular architecture and reusability functionalities of "JTrack" enable the researchers to easily add or remove their required features from a list of collected data.

In the second manuscript (68), we compared "JTrack" with two stationary gait analysis systems (a force plate and a motion capture system). We found a high level of agreement between data collected from single accelerometer sensors using the "JTrack" application and stationary analysis systems, demonstrating the utility of smartphones in future clinical studies of gait. We have shown that features such as stride time and cadence variables, clinically meaningful outcomes of locomotor tasks, can be accurately derived from smartphones and used in a normal-gait analysis. However, we came to the conclusion that measuring in particular stance variables has limitations, such as the need for initial calibration or a device's fixed orientation and position in the reference recording.

The DBs are inherently prone to different sources of variation. Recent studies that have demonstrated the feasibility of using DBs in PD have mainly focused on in-laboratory settings, therefore, validation and reliability studies are still lacking for home environments. In the third manuscript (69), we addressed this issue by evaluating the test-retest reliability and longitudinal stability of DBs for PD as measured in a large-scale m-Power study using a self-administered setting (66). We observed a significant change in the longitudinal performance of most DBs, resulting in poor differentiation between PD and HC. We attributed this finding to the presence of differential learning and variation in motivation between the two groups. Motivation and learning effects are two barely studied sources of confounds that impact the longitudinal stability of DBs. Although learning has been identified as a crucial factor that can influence a participant's performance over time (70–72), the effect of motivation on DBs' reliability remains understudied. Our study found evidence for both aspects, highlighting the need for an investigation that explores the effect of motivation alteration and how it differs from learning effects. We also noted that a decrease in motivation could impact adherence to the study, as evidenced by the limited records of most participants during the course of the study in the m-Power dataset.

Initial investigation of the m-Power dataset indicates that using the self-administration enrolment method may lead to several biases, such as the recruitment of younger and healthier subjects. This may be explained by the recruitment restrictions of m-power (owning an iPhone, speaking English, and residing in the United States) targeting a young and healthy population with a higher smartphone adaptation rate (73). Therefore, demographic factors such as age, technological savvy (primarily among the elderly), and socioeconomic status may influence the profile of participation in smartphone-based remote monitoring programs. Hence, it is crucial to consider selection bias in such research.

Additionally, the data collected from participants' self-reports may introduce several biases stemming from incorrect data entry. The ability to follow or understand instructions required for completing tasks is another concept that can be a source of bias. This is more pronounced for complex tasks such as gait and balance. Task instructions such as "take ten steps and turn around" and "start walking for ten steps after you hear a beep sound" may be misunderstood or entirely ignored by the participants. This effect was observed in the m-Power dataset, where many participants had shorter signal lengths or held devices in the wrong positions. In addition, the results of the third manuscript show a smaller effect size and poor reliability for these tasks. (6,74). It suggests that additional supplemental materials, such as audio counselling or video tutorials, during or before such tasks could help participants complete them.

The real-time and longitudinal nature of remote monitoring approaches makes them an excellent complement to patient recall for the estimation of treatment effect evaluation, with the potential to be utilised in personalised and precision medicine. However, our results indicate that the majority of extracted DBs were not sensitive enough to different treatment conditions. One potential explanation is the presence of confounding factors such as the absence of accurate treatment reporting time, treatment type, dosage, and disease severity that must all be considered. Thus, there is a need for more sensitive DBs in a remote and self-administered context as well as carefully constructed settings and evaluation procedures in this response.

Upon further inspection of manuscripts published using the m-Power data set, it became clear that there was a vast divergence of outcomes from different groups. One of the most potential explanations for this disparity is the lack of a standard pipeline for analysis, standardised benchmarks, performance measures, and neglect of confounding factors. Therefore, in the fourth manuscript (75), we investigated these confounding factors and their influence on the performance of machine learning methods.

Different machine learning algorithms search for particular trends and patterns in the data, and it's possible that a single algorithm may not be the optimal solution for all datasets or use cases. Thus, comprehensive experiments and the assessment of multiple machine-learning algorithms are required. We show this principle in the fourth paper by comparing the performances of different machine-learning algorithms. We also show that different sources of confounding factors can lead to over or underfitting results, which is consistent with previous research (58,76).

Considering that PD is a complex disease affecting many health domains, using a single task or a limited number of features could be another reason for the discrepancy in previous studies. Therefore, considering multiple aspects such as sleep behaviour, genotyping, and mood tracking to extract more reliable DBs may increase the performance of machine learning methods in differentiating PD from non-PD groups (15).

The large volume of data collected from smartphones typically consists of repeated measurements from a participant. The assignment of repeated measures from the same individual (record-wise) to training and test sets is a common practice in machine learning methods. However, this potentially misleads the algorithm into learning the unique feature associated with a subject resulting in a very optimistic performance. Thus to avoid this, it is recommended to use subject-wise data separation rather than record-wise. (65,77). Finally, the outcome of the fourth manuscript suggests that the lack of considering the above-mentioned factors could lead to a lack of reproducibility and generalizability in the results.

Conclusion and future direction

In conclusion, in this thesis, I have highlighted the potential of using DBs in clinical studies, enabling a diverse range of research and clinical applications such as disease monitoring and diagnosis. I also pointed out that developing successful assessment tools require close collaboration between medical professionals, analytical experts, computer engineers, and legal experts, which cannot be achieved without a broad management operation. Therefore, the availability of an open-source and modular platform enables clinicians and researchers to incorporate such a tool into their studies quickly. In addition, concerns such as privacy, performance, security, and data ownership are among the fundamental factors that have the potential to influence this partnership. Therefore, tools should be most sensitive to these issues.

I also showed that the reliability (test-retest) and longitudinal stability of DBs extracted in an unsupervised manner is a major challenge regarding their clinical use. The complexity of self-administered and at-home protocols, along with factors such as environmental changes, loss of motivation, age and gender differences, and the influence of comorbidities, are factors that can cause DBs to become unstable over time, which have not been well addressed in previous studies. Therefore, more comprehensive mechanisms for the development and analysis of DBs collected under these conditions, as well as the development of more reliable DBs, need to be addressed in future studies. Also, study participation and adherence are other critical challenges associated with uncontrolled protocols, improved study protocols and using intelligent notification combined with passive monitoring methods may improve this issue.

Finally, before the widespread deployment of DBs technologies in clinical practice, remote assessment platforms need to be upgraded to be compatible with various operating systems and devices. Integrating innovative embedded sensors and wearable devices should also be implemented in future updates of DB platforms. In addition, combining behavioural data with sensor-based data is among the things that may reveal more information about the disease.

Reference

1. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA. 2017 Oct 3;318(13):1215–6.
2. Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord. 2018 Aug;33(8):1287–97.
3. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. npj Digital Med. 2019 Mar 11;2(1).
4. Patel S, Lorincz K, Hughes R, Huggins N, Growdon J, Standaert D, et al. Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. IEEE Trans Inf Technol Biomed. 2009 Nov;13(6):864–73.
5. Sejdić E, Lowry KA, Bellanca J, Redfern MS, Brach JS. A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains. IEEE Trans Neural Syst Rehabil Eng. 2014 May;22(3):603–12.
6. Omberg L, Chaibub Neto E, Perumal TM, Pratap A, Tediarjo A, Adams J, et al. Remote smartphone monitoring of Parkinson's disease and individual response to therapy. Nat Biotechnol. 2022 Apr;40(4):480–7.
7. Zhan A. High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection. CoRR. 2016;
8. Prince J, Arora S, de Vos M. Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. Physiol Meas. 2018 Apr 26;39(4):044005.
9. Hossain SM, Hnat T, Saleheen N, Nasrin NJ, Noor J, Ho B-J, et al. mCerebrum: A Mobile Sensing Software Platform for Development and Validation of Digital Biomarkers and Interventions. Proc Int Conf Embed Netw Sens Syst. 2017 Nov;2017.
10. Ranjan Y, Rashid Z, Stewart C, Kerz M, Begale M, Verbeeck D, et al. RADAR-base: An Open Source mHealth Platform for Collecting, Monitoring and Analyzing data Using Sensors, Wearables, and Mobile Devices (Preprint). 2018 Aug 29;
11. Ferreira D, Kostakos V, Dey AK. AWARE: mobile context instrumentation framework. Front ICT. 2015 Apr 20;2(6):1–9.
12. Torous J, Kiang MV, Lorme J, Onnela J-P. New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health. 2016 May 5;3(2):e16.
13. Wang X, Vouk N, Heaukulani C, Buddhika T, Martanto W, Lee J, et al. HOPES: an integrative digital phenotyping platform for data collection, monitoring, and machine learning. J Med Internet Res. 2021 Mar 15;23(3):e23984.
14. Baran SW, Bratcher N, Dennis J, Gaburro S, Karlsson EM, Maguire S, et al. Emerging role of translational digital biomarkers within home cage monitoring technologies in preclinical drug discovery and development. Front Behav Neurosci. 2021;15:758274.
15. Deng K, Li Y, Zhang H, Wang J, Albin RL, Guan Y. Heterogeneous digital biomarker integration out-performs patient self-reports in predicting Parkinson's disease. Commun Biol. 2022 Jan 17;5(1):58.
16. Stephenson D, Badawy R, Mathur S, Tome M, Rochester L. Digital progression biomarkers as novel endpoints in clinical trials: A multistakeholder perspective. J Parkinsons Dis. 2021 Feb 11;

17. Estimation of the 2020 Global Population of Parkinson's Disease (PD) - MDS Abstracts [Internet]. [cited 2022 Jul 4]. Available from: https://www.mdsabstracts.org/abstract/estimation-of-the-2020-global-population-of-parkinsons-disease-pd/
18. Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry. 2008 Apr;79(4):368–76.
19. Bang Y, Lim J, Choi HJ. Recent advances in the pathology of prodromal non-motor symptoms olfactory deficit and depression in Parkinson's disease: clues to early diagnosis and effective treatment. Arch Pharm Res. 2021 Jun 19;44(6):588–604.
20. Zhu B, Yin D, Zhao H, Zhang L. The immunology of Parkinson's disease. Semin Immunopathol. 2022 Jun 8;
21. Obeso JA, Rodriguez-Oroz MC, Goetz CG, Marin C, Kordower JH, Rodriguez M, et al. Missing pieces in the Parkinson's disease puzzle. Nat Med. 2010 Jun;16(6):653–61.
22. Venda LL, Cragg SJ, Buchman VL, Wade-Martins R. α-Synuclein and dopamine at the crossroads of Parkinson's disease. Trends Neurosci. 2010 Dec;33(12):559–68.
23. Rouaud T, Corbillé AG, Leclair-Visonneau L, de Guilhem de Lataillade A, Lionnet A, Preterre C, et al. Pathophysiology of Parkinson's disease: Mitochondria, alpha-synuclein and much more…. Rev Neurol. 2021 Mar;177(3):260–71.
24. Farrer M, Kachergus J, Forno L, Lincoln S, Wang D-S, Hulihan M, et al. Comparison of kindreds with parkinsonism and alpha-synuclein genomic multiplications. Ann Neurol. 2004 Feb;55(2):174–9.
25. Dauer W, Przedborski S. Parkinson's disease: mechanisms and models. Neuron. 2003 Sep 11;39(6):889–909.
26. Gasser T. Update on the genetics of Parkinson's disease. Mov Disord. 2007 Sep;22 Suppl 17:S343-50.
27. Hindle JV. Ageing, neurodegeneration and Parkinson's disease. Age Ageing. 2010 Mar;39(2):156–61.
28. Lema Tomé CM, Tyson T, Rey NL, Grathwohl S, Britschgi M, Brundin P. Inflammation and α-synuclein's prion-like behavior in Parkinson's disease--is there a link? Mol Neurobiol. 2013 Apr;47(2):561–74.
29. Yang W, Hamilton JL, Kopil C, Beck JC, Tanner CM, Albin RL, et al. Current and projected future economic burden of Parkinson's disease in the U.S. npj Parkinsons Disease. 2020 Jul 9;6:15.
30. Dorsey ER, Papapetropoulos S, Xiong M, Kieburtz K. The first frontier: digital biomarkers for neurodegenerative disorders. Digit Biomark. 2017 Dec;1(1):6–13.
31. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, 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 Nov 15;23(15):2129–70.
32. Zhan A, Little MA, Harris DA, Abiola SO, Dorsey ER, Saria S, et al. High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection. arXiv. 2016 Jan 5;
33. Rovini E, Maremmani C, Cavallo F. How wearable sensors can support parkinson's disease diagnosis and treatment: A systematic review. Front Neurosci. 2017 Oct 6;11:555.
34. Solhan MB, Trull TJ, Jahng S, Wood PK. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. Psychol Assess. 2009 Sep;21(3):425–36.

35. Adler CH, Beach TG, Hentz JG, Shill HA, Caviness JN, Driver-Dunckley E, et al. Low clinical diagnostic accuracy of early vs advanced Parkinson disease: clinicopathologic study. Neurology. 2014 Jul 29;83(5):406–12.
36. Hausdorff JM. Gait dynamics in Parkinson's disease: common and distinct behavior among stride length, gait variability, and fractal-like scaling. Chaos. 2009 Jun;19(2):026113.
37. Schlachetzki JCM, Barth J, Marxreiter F, Gossler J, Kohl Z, Reinfelder S, et al. Wearable sensors objectively measure gait parameters in Parkinson's disease. PLoS One. 2017 Oct 11;12(10):e0183989.
38. Horak FB, Mancini M. Objective biomarkers of balance and gait for Parkinson's disease using body-worn sensors. Mov Disord. 2013 Sep 15;28(11):1544–51.
39. Rampp A, Barth J, Schülein S, Gaßmann K-G, Klucken J, Eskofier BM. Inertial sensor-based stride parameter calculation from gait sequences in geriatric patients. IEEE Trans Biomed Eng. 2015 Apr;62(4):1089–97.
40. Cho Y-S, Jang S-H, Cho J-S, Kim M-J, Lee HD, Lee SY, et al. Evaluation of Validity and Reliability of Inertial Measurement Unit-Based Gait Analysis Systems. Ann Rehabil Med. 2018 Dec 28;42(6):872–83.
41. Cancela J, Pastorino M, Arredondo MT, Nikita KS, Villagra F, Pastor MA. Feasibility study of a wearable system based on a wireless body area network for gait assessment in Parkinson's disease patients. Sensors (Basel). 2014 Mar 7;14(3):4618–33.
42. San-Segundo R, Torres-Sánchez R, Hodgins J, De la Torre F. Increasing robustness in the detection of freezing of gait in parkinson's disease. Electronics. 2019 Jan 22;8(2):119.
43. Bächlin M, Plotnik M, Roggen D, Maidan I, Hausdorff JM, Giladi N, et al. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. IEEE Trans Inf Technol Biomed. 2010 Mar;14(2):436–46.
44. Blin O, Ferrandez AM, Serratrice G. Quantitative analysis of gait in Parkinson patients: increased variability of stride length. J Neurol Sci. 1990 Aug;98(1):91–7.
45. Prieto TE, Myklebust JB, Hoffmann RG, Lovett EG, Myklebust BM. Measures of postural steadiness: differences between healthy young and elderly adults. IEEE Trans Biomed Eng. 1996 Sep;43(9):956–66.
46. 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. 2012;15(9):899–910.
47. Palakurthi B, Burugupally SP. Postural instability in parkinson's disease: A review. Brain Sci. 2019 Sep 18;9(9).
48. Foreman KB, Sondrup S, Dromey C, Jarvis E, Nissen S, Dibble LE. The effects of practice on the concurrent performance of a speech and postural task in persons with Parkinson disease and healthy controls. Parkinsons Dis. 2013 Jun 11;2013:987621.
49. Mahadevan N, Demanuele C, Zhang H, Volfson D, Ho B, Erb MK, et al. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. npj Digital Med. 2020 Jan 15;3:5.
50. Cancela J, Pastorino M, Tzallas AT, Tsipouras MG, Rigas G, Arredondo MT, et al. Wearability assessment of a wearable system for Parkinson's disease remote monitoring based on a body area network of sensors. Sensors (Basel). 2014 Sep 16;14(9):17235–55.
51. 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. 2011 May;15(3):481–90.

52. 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. 2017;9:11–7.

53. Ma A, Lau KK, Thyagarajan D. Voice changes in Parkinson's disease: What are they telling us? J Clin Neurosci. 2020 Feb;72:1–7.

54. 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. 2020 Jun 1;70(11):393–405.

55. Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform. 2020 Apr;104:103362.

56. Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online. 2007 Jun 26;6:23.

57. 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 Dec 10;2013:389195.

58. Schwab P, Karlen W. PhoneMD: Learning to Diagnose Parkinson's Disease from Smartphone Data. AAAI. 2019 Jul 17;33:1118–25.

59. Arora S, Venkataraman V, Zhan A, Donohue S, Biglan KM, Dorsey ER, et al. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. Parkinsonism Relat Disord. 2015 Jun;21(6):650–3.

60. Prince J, Andreotti F, De Vos M. Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-Wise Missing Data. IEEE Trans Biomed Eng. 2019;66(5):1402–11.

61. Tsanas A, Little MA, McSharry PE, Ramig LO. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. J R Soc Interface. 2011 Jun 6;8(59):842–55.

62. Cheng W-Y, Scotland A, Lipsmeier F, Kilchenmann T, Jin L, Schjodt-Eriksen J, et al. Human Activity Recognition from Sensor-Based Large-Scale Continuous Monitoring of Parkinson's Disease Patients. 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). IEEE; 2017. p. 249–50.

63. Kabelac Z, Tarolli CG, Snyder C, Feldman B, Glidden A, Hsu C-Y, et al. Passive monitoring at home: A pilot study in parkinson disease. Digit Biomark. 2019 Apr 30;3(1):22–30.

64. Roussos G, Herrero TR, Hill DL, Dowling AV, L T M Müller M, Evers LJW, et al. Identifying and characterising sources of variability in digital outcome measures in Parkinson's disease. npj Digital Med. 2022 Jul 15;5(1):93.

65. Chaibub Neto E, Pratap A, Perumal TM, Tummalacherla M, Snyder P, Bot BM, et al. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. npj Digital Med. 2019 Oct 11;2:99.

66. Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data. 2016 Mar 3;3:160011.

67. Espay AJ, Bonato P, Nahab FB, Maetzler W, Dean JM, Klucken J, et al. Technology in Parkinson's disease: Challenges and opportunities. Mov Disord. 2016 Sep;31(9):1272–82.

68. Rentz C, Far MS, Boltes M, Schnitzler A, Amunts K, Dukart J, et al. System Comparison for Gait and Balance Monitoring Used for the Evaluation of a Home-Based Training. Sensors (Basel). 2022 Jun 30;22(13).

69. Sahandi Far M, Eickhoff SB, Goni M, Dukart J. Exploring Test-Retest Reliability and Longitudinal Stability of Digital Biomarkers for Parkinson Disease in the m-Power Data Set: Cohort Study. J Med Internet Res. 2021 Sep 13;23(9):e26608.

70. Olson M, Lockhart TE, 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. 2019 Feb 7;10:62.

71. Krebs HI, Hogan N, Hening W, Adamovich SV, Poizner H. Procedural motor learning in Parkinson's disease. Exp Brain Res. 2001 Dec;141(4):425–37.

72. Behrman AL, Cauraugh JH, Light KE. Practice as an intervention to improve speeded motor performance and motor learning in Parkinson's disease. J Neurol Sci. 2000 Mar;174(2):127–36.

73. Demographics of Mobile Device Ownership and Adoption in the United States | Pew Research Center [Internet]. [cited 2022 Dec 14]. Available from: https://www.pewresearch.org/internet/fact-sheet/mobile/

74. Badawy R, Raykov YP, Evers LJW, Bloem BR, Faber MJ, Zhan A, et al. Automated quality control for sensor based symptom measurement performed outside the lab. Sensors (Basel). 2018 Apr 16;18(4).

75. Goñi M, Eickhoff S, Far MS, Patil K, Dukart J. Limited diagnostic accuracy of smartphone-based digital biomarkers for Parkinson's disease in a remotely-administered setting. medRxiv. 2021 Jan 15;

76. Neto EC, Pratap A, Perumal TM, Tummalacherla M, Bot BM, Mangravite L, et al. Using permutations to assess confounding in machine learning applications for digital health. arXiv. 2018;

77. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. Gigascience. 2017 May 1;6(5):1–9.

Acknowledgement

This dissertation could not have been written without the assistance of several people to whom I am greatly indebted.

I would like to express my gratitude to my direct supervisor PD. Dr Juergen Dukart for his guidance, support, and encouragement throughout my study. I would like to express my appreciation to Prof. Dr Simon B. Eickhoff for all his helpful and kind supervision. I greatly appreciate Prof. Dr Oliver Kuß for his support and guidance.

I would also like to thank all the co-authors of my papers, without whom this effort could not have been completed. In addition, I would like to thank the committee members of my thesis defence for their insightful feedback.

I would like to acknowledge the funding sources for my projects: The Human Brain Project, supported by the European Union's Horizon 2020 Framework Program for Research and Innovation under Specific Grant Agreement No. 785907. (Human Brain Project SGA2).

Finally, I also thank the Research Centre Jülich for providing all financial and computational resources and all Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7) members for all the support and good time we shared.