frontiers | Research Topics

Digital remote patient monitoring in neurodegenerative diseases

Edited by
Graham Jones, Amit Khanna and Diane Stephenson

Published in
Frontiers in Digital Health

ISSN 1664-8714
ISBN 978-2-8325-7200-9
DOI 10.3389/978-2-8325-7200-9

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Digital remote patient monitoring in neurodegenerative diseases

Topic editors

Graham Jones — Tufts Medical Center, United States

Amit Khanna — Novartis, Switzerland

Diane Stephenson — Critical Path Institute, United States

Citation

Jones, G., Khanna, A., Stephenson, D., eds. (2025). *Digital remote patient monitoring in neurodegenerative diseases*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-7200-9

Table of contents

frontiers | Frontiers in Digital Health

TYPE Editorial
PUBLISHED 11 November 2025
DOI 10.3389/fdgth.2025.1725819

OPEN ACCESS

EDITED AND REVIEWED BY
Toshiyo Tamura,
Waseda University, Japan

*CORRESPONDENCE
Amit Khanna
✉ amit.khanna@novartis.com
Diane Stephenson
✉ dstephenson@cpath.org
Graham Jones
✉ graham.jones@tufts.edu

RECEIVED 15 October 2025
ACCEPTED 20 October 2025
PUBLISHED 11 November 2025

CITATION
Khanna A, Stephenson D and Jones G (2025) Editorial: Digital remote patient monitoring in neurodegenerative diseases.
Front. Digit. Health 7:1725819.
doi: 10.3389/fdgth.2025.1725819

Editorial: Digital remote patient monitoring in neurodegenerative diseases

Amit Khanna[1]*, Diane Stephenson[2]* and Graham Jones[3]*

[1]Novartis AG, Basel, Switzerland, [2]Critical Path Institute, Tucson, AZ, United States, [3]Tufts Medical Center, Boston, MA, United States

KEYWORDS

remote monitoring, sensors, disease progression, composite digital biomarkers, speech, ocular, regulatory pathways, real world evidence

Editorial on the Research Topic
Digital remote patient monitoring in neurodegenerative diseases

Remote monitoring is gaining prominence in patient care enabling the recording of repeat measurements outside of scheduled clinical visits. In neurodegenerative diseases this is of added significance as changes that take place are often subtle, span multiple domains (motor, cognition, sleep, speech, etc.) and new symptoms develop over several years. This special volume integrates perspectives from experts and thought leaders in the academy, pharmaceutical industry, and research foundations who highlight significant new developments.

One of the important elements for adoption of remote monitoring devices in clinical studies is patient acceptance and to adhere over sustained periods necessary to derive clinically meaningful data. In a review of wearable device adoption rates, Hirczy et al. conducted online surveys to identify barriers to uptake among Parkinson's disease patients. Surprisingly, among US based patients although greater than 90% of respondents were interested in new technologies only 24% were using consumer devices for disease management and only 8% with medical grade wearables.

Similarly, Kangarloo et al. report patient experiences with body-worn sensors used in clinic and a mobile application used at-home from the WATCH-PD study. This observational, 12 month study focused on disease progression in early Parkinson's Disease among 82 participants with PD and 50 control participants. Results demonstrated that participants had generally positive views on comfort and use of the technologies throughout the study duration regardless of group. Significantly, device proficiency and acceptability in people with early stage PD did not differ from neurologically healthy older adults, providing impetus for future clinical studies.

Careful study design is paramount when implementing new technologies in clinical settings including assessing the reliability of the data captured. Lavine et al. examine the test-retest reliability of accelerometry derived data from at-home studies. Using raw data derived from triaxial accelerometry involving 21 PD patients and 23 controls they applied linear mixed models to determine the identity of drug treatment effects. They conclude that at-home measures have favorable reliability profiles as more data points can be gathered, and the reduction in sample size needed to detect progression presents clear justification for their deployment in future studies.

The design of long term studies of disease tracking will likely require development of innovative computational approaches to data capture and interpretation. Zhai et al. present a new machine learning framework to construct composite digital biomarkers for progression tracking. The framework was applied to data collected from an observational PD study involving movement measurements captured using the Opal TM sensor combined with MDS-UPDRS Part III scores. The composite digital measure exhibited a smoother and more significant increasing trend over time with less variability, and ability to classify between *de novo* PD and healthy controls.

Although a majority of studies have focused on movement and motion tracking, there are a number of exciting developments on the horizon with alternative measures. Speech and acoustic signals are a potentially very rich source of clinical information in neurological diseases and Tröger et al. highlight recent findings on speech intelligibility. They describe a digital measure for speech intelligibility which was deployed on datasets from patients suffering from Dysarthria, a motor speech disorder associated with Parkinson's Disease (PD), progressive supranuclear palsy (PSP), Huntington's Disease (HD) and amyotrophic lateral sclerosis (ALS). The score, derived from automatic speech recognition (ASR) systems, showed good to excellent inter-rater reliability and significant differences in intelligibility scores between pathological groups and healthy controls.

Ocular analysis is another area of promise and Band et al. provide a timely overview of the study of eye movement abnormalities to indicate neurological condition severity and distinguish disease phenotypes. Recent strides in imaging sensors and computational power have resulted in a surge in the development of technologies facilitating the extraction and analysis of eye movements to assess neurodegenerative diseases. Their review provides an overview of these advancements, their potential to offer patient-friendly assessments and explores current trends and future directions in this exciting field.

Other approaches are being developed with the similar goal of detecting diseased states at population level using low patient burden technologies. Jiang et al. reflect on studies in Canada where automated facial expression analysis (AFEA) was compared to standard measures such as electroencephalogram (EEG) technologies and heart rate variability (HRV). The case for development of composite measures of cognitive decline based on AFEA is presented, and its utility in remote deployment using contactless data capture supported by potential economic benefits through the national healthcare system.

Advancing digital remote monitoring technologies for drug development studies requires careful approach to study design and ultimately alignment with prevailing regulatory guidance. In a timely overview the role of the Critical Path Institute is highlighted, bridging key interfaces between the health authorities, pharmaceutical industry sponsors, patient advocates, and the clinical research community Stephenson et al. Progress made through the Critical Path for Parkinson's Consortium's (CPP) Digital Drug Development tool (3DT) serves to showcase their approach. The initiative has helped accelerate the regulatory maturity of several key digital health technology measures, and advanced thinking on approaches to clinical trial design, data acquisition and the use of AI methodologies to extract critical features.

A tenet in regulatory guidance for remote patient assessment is the need to focus on activities of daily living (ADL) and real world evidence. An emerging trend for patient monitoring is the development of smart home environments, with sensors and devices located strategically to capture key health related data. Grammatikopoulou et al. report findings on the assessment of ADLs in subjects at the CERTH-IT simulated Smart Home. Sensor data was used to track activity as subjects (controls and groups suffering from cognitive decline) conducted various tasks and operations. Differentiation between controls and other groups was attainable and valuable feedback obtained to refine the approach for wider deployment.

These are exciting times for the deployment of patient monitoring technologies in neurodegenerative diseases. Progress highlighted by these leaders is having demonstrable impact on moving the field forward. We hope this inspires others to innovate, challenge hypotheses, and develop practical solutions to advance new treatment options and ultimately influence patient care. Clearly, there is more to come.

Sincerely,
Amit Khanna, Diane Stephenson, Graham Jones

Author contributions

AK: Conceptualization, Writing – original draft, Writing – review & editing. DS: Conceptualization, Writing – original draft, Writing – review & editing. GJ: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The 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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

frontiers | Frontiers in Aging Neuroscience

TYPE Original Research
PUBLISHED 28 March 2024
DOI 10.3389/fnagi.2024.1375131

OPEN ACCESS

EDITED BY
Michael Ewers,
Ludwig Maximilian University of Munich, Germany

REVIEWED BY
Graham Jones,
Tufts Medical Center, United States
Katharina Geschke,
Johannes Gutenberg University Mainz, Germany

*CORRESPONDENCE
Margarita Grammatikopoulou
✉ marggram@iti.gr

RECEIVED 23 January 2024
ACCEPTED 06 March 2024
PUBLISHED 28 March 2024

CITATION
Grammatikopoulou M, Lazarou I, Alepopoulos V, Mpaltadoros L, Oikonomou VP, Stavropoulos TG, Nikolopoulos S, Kompatsiaris I and Tsolaki M (2024) Assessing the cognitive decline of people in the spectrum of AD by monitoring their activities of daily living in an IoT-enabled smart home environment: a cross-sectional pilot study. *Front. Aging Neurosci.* 16:1375131. doi: 10.3389/fnagi.2024.1375131

Assessing the cognitive decline of people in the spectrum of AD by monitoring their activities of daily living in an IoT-enabled smart home environment: a cross-sectional pilot study

Margarita Grammatikopoulou[1]*, Ioulietta Lazarou[1], Vasilis Alepopoulos[1], Lampros Mpaltadoros[1], Vangelis P. Oikonomou[1], Thanos G. Stavropoulos[1], Spiros Nikolopoulos[1], Ioannis Kompatsiaris[1], Magda Tsolaki[2,3,4] and on behalf of RADAR-AD

[1]Information Technologies Institute, Centre for Research and Technology Hellas (CERTH-ITI), Thessaloniki, Greece, [2]1st Department of Neurology, G.H. "AHEPA", School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece, [3]Greek Association of Alzheimer's Disease and Related Disorders (GAADRD), Thessaloniki, Greece, [4]Laboratory of Neurodegenerative Diseases, Center for Interdisciplinary Research and Innovation (CIRI - AUTh), Balkan Center, Buildings A & B, Aristotle University of Thessaloniki, Thessaloniki, Greece

Introduction: Assessing functional decline related to activities of daily living (ADLs) is deemed significant for the early diagnosis of dementia. As current assessment methods for ADLs often lack the ability to capture subtle changes, technology-based approaches are perceived as advantageous. Specifically, digital biomarkers are emerging, offering a promising avenue for research, as they allow unobtrusive and objective monitoring.

Methods: A study was conducted with the involvement of 36 participants assigned to three known groups (Healthy Controls, participants with Subjective Cognitive Decline and participants with Mild Cognitive Impairment). Participants visited the CERTH-IT Smart Home, an environment that simulates a fully functional residence, and were asked to follow a protocol describing different ADL Tasks (namely Task 1 – Meal, Task 2 – Beverage and Task 3 – Snack Preparation). By utilizing data from fixed in-home sensors installed in the Smart Home, the identification of the performed Tasks and their derived features was explored through the developed CARL platform. Furthermore, differences between groups were investigated. Finally, overall feasibility and study satisfaction were evaluated.

Results: The composition of the ADLs was attainable, and differentiation among the HC group compared to the SCD and the MCI groups considering the feature "Activity Duration" in Task 1 – Meal Preparation was possible, while no difference could be noted between the SCD and the MCI groups.

Discussion: This ecologically valid study was determined as feasible, with participants expressing positive feedback. The findings additionally reinforce the interest and need to include people in preclinical stages of dementia in research to further evolve and develop clinically relevant digital biomarkers.

KEYWORDS

Alzheimer's disease, healthy controls, subjective cognitive decline, mild cognitive impairment, smart home, sensor technology, activities of daily living

1 Introduction

According to the World Health Organisation (WHO), there are currently over 55 million people living with dementia (PwD) globally (World Health Organisation Dementia Key Facts, 2022). The sharp increase in dementia cases is likely to have significant consequences for healthcare providers, caregivers, and the economy (Aranda et al., 2021). For this, research has focused on the early detection of dementia with the primary objective to intervene before symptoms worsen and lead to loss of independence and greater need for care (Rasmussen and Langerman, 2019).

This is further supported by the fact that search for effective treatments of AD has led to the first disease-modifying therapies (Lecanemab and Aducanumab). These treatments have been approved by the FDA as well as in Japan and are being considered by the EMA (European Medicines Agency, 2023). Furthermore, 141 drugs are currently being tested in clinical trials for the treatment of AD, 80% of which aim to slow disease progression (Cummings et al., 2023).

The need to identify people at the pre-symptomatic stage becomes eminent, as the recently developed therapeutic agents exhibit their greatest potential in early AD (Cummings et al., 2023; van Dyck et al., 2023). Additionally, lifestyle and other non-pharmacological interventions (e.g., the multidomain FINGER intervention (Ngandu et al., 2022)), show promising results in preventing symptom progression when applied timely, before the onset of dementia.

An early sign of dementia is functional deterioration expressed often through difficulties in performing Activities of Daily Living (ADLs), as an association has been found to exist between ADL deficits and cognitive functioning (Bangen et al., 2010; Jekel et al., 2015). Current approaches for assessing ADLs to determine functional decline involve traditional pen and paper methods. As these rely on informant input and are often not sensitive enough to capture subtle changes, there is further space for improvement and development of complementary measures (Sikkes et al., 2009).

Shifting focus to unobtrusive, passive, objective monitoring approaches, digital biomarkers have emerged showing promising potential (Anna-Katharine et al., 2023). In a general sense, using technology-based approaches to evaluate ADLs in older adults is a promising area of research with several advantages over traditional cognitive assessment methods. However, a major drawback of these tools is that they may require prolonged use to detect subtle ADL differences that indicate cognitive decline. Nevertheless, the obtained information from digital biomarkers, reflect real-life conditions, while eliminating reporting bias. They can be derived from passive sensors, wearables, purposive technological solutions (e.g., games) and other technological solutions (e.g., assessment of computer mouse movements, identify if pill box used) (Piau et al., 2019). Digital biomarkers can be used to assess walking and sleep patterns, physical activity and also, ADLs. They represent a valuable method, as they comprise sensitive and precise measures that can detect subtle changes. This makes them suitable in assessing deterioration in function that can occur at an early, preclinical stage.

A plethora of sensors has been used and deployed in the context of Smart Homes (SH) (in the sense of controlled research environments, care homes or participants' homes where the sensors are being installed) allowing for remote in-home sensing and remote ADL monitoring (Garcia-Constantino et al., 2021; Moyle et al., 2021). There are many opportunities for the use of various SH technologies in community-dwelling PwD, ranging from diagnostic assessment to long-term and personalized care management. As a result, many individual studies have been conducted on the development and use of SH technologies in older populations (Ma et al., 2023; Yu et al., 2023). Such technologies are being investigated for use in a wide range of applications and contexts. These can vary from home based monitoring, personalized care, quality of life improvement, to independent living, observation and prediction of the actions of a person, caregiver burden reduction, intervention and disease progression monitoring, and also identification of emergency situations (Amiribesheli and Bouchachia, 2018; Ault et al., 2020; Han et al., 2022; Miller et al., 2022). Furthermore, there is a growing interest in the use of using digital biomarkers assessing ADLs, as reliable proxies for screening participants for clinical trials or as secondary endpoints (Atkins et al., 2015; Gold et al., 2018).

The use of sensor technology to identify cognitive decline through observing ADL performance is not a novel concept. Even so, the field of exploring methods and developing digital biomarkers to quantify and compare ADL performance is still in its infancy.

1.1 Aim of the present work

This work has been conducted in the context of RADAR-AD,[1] an EU-funded project that explores the potential of mobile and digital technologies to improve the assessment of Alzheimer's Disease (AD) (Owens et al., 2020; Muurling et al., 2021). In particular, the main motivation in one of RADAR-AD's sub-studies was to explore whether the identification and monitoring of ADLs was achievable, utilizing data collected from in-home sensors in a Smart Home environment. Furthermore, it was investigated if the identified ADLs can provide clinically meaningful insights regarding the preclinical stages of AD. Additionally, technology acceptance and the overall feasibility of the study was assessed.

In detail, we assessed a number of people at preclinical and prodromal stages of AD, namely, the Subjective Cognitive Decline (SCD) stage, and the Mild Cognitive Impairment (MCI) stage (Dubois et al., 2016), that were evaluated against healthy control (HC) participants in terms of their performance during the execution of particular ADLs. Their performance was monitored through the data collected by a set of commercially available fixed in-home sensors[2] installed in CERTH-ITI's Smart Home.[3] The sensor data were collected, processed and visualized using a platform developed by our research team (Mpaltadoros et al., 2021). First insights could be gained, regarding the effectiveness of remotely monitoring ADLs and their potential to offer quantifiable metrics for discriminating between the different stages of cognitive impairment. Furthermore, all participants filled a detailed questionnaire assessing overall study satisfaction while staying at CERTH-ITI's SH, evaluating the presented sensor technologies. The study pipeline is given in Figure 1.

1 https://www.radar-ad.org/

2 FIBARO sensors: https://www.fibaro.com/en/products/all-domotica-devices/

3 https://smarthome.iti.gr/

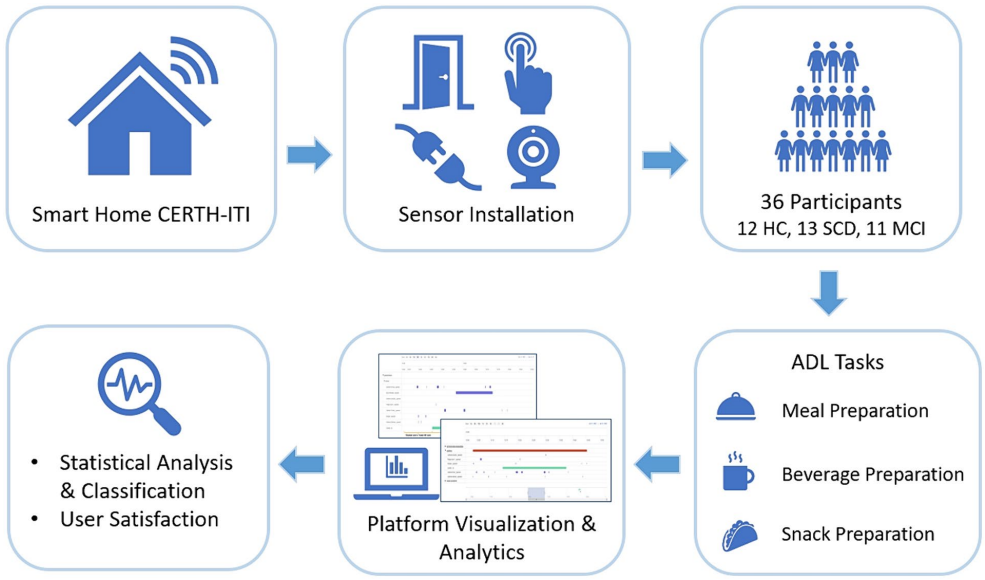


FIGURE 1
Study pipeline.

2 Materials and methods

2.1 Study protocol

2.1.1 Participants

Participants were recruited from the Greek Association of Alzheimer's Disease and Related Disorders (GAADRD)[4] and a wide community audience. The study was carried out in accordance with the Declaration of Helsinki and received approval by the Ethics Committee of CERTH (ETH.COM 54/17-06-2020) and the Scientific and Ethics Committee of GAADRD (242/2022 AI_07/10/2021), while a written informed consent was obtained from all participants prior to their participation in the study. The Information Forms used to debrief the participants were prepared according to ICH-GCP requirements and data protection regulations [European Medicines Agency (EMA), 2016].

The diagnosis of HC, SCD and MCI was set by a neuropsychiatrist, specialised in dementia, according to the structural magnetic resonance imaging (MRI), medical history, neuropsychological tests and neurological examination. The MCI group fulfilled the Petersen criteria (Petersen et al., 2009) and it is noted that all MCI cases were of the amnestic subtype. The SCD group met IWG-2 Guidelines (Dubois et al., 2014) as well as the SCD-I Working Group instructions (Molinuevo et al., 2017). Regarding the SCD and MCI groups, we excluded participants with confounding factors based on blood tests (hormonal disorders, vitamin deficiency etc.), while structural MRI scans were done for participants in both groups (vascular/demyelinating lesions, tumours, anatomical variations etc.). Additional inclusion criteria for the SCD and HC participants included having a normal general medical, neurological and neuropsychological examination. Exclusion criteria comprised severe psychiatric, physical or other neurological disorder, illness or any other somatic disorder, which may cause cognitive impairment. Additionally, it is noted that as the study protocol included an EEG based action [explored in Ioulietta et al., (2023)], left-handedness constituted an exclusion criterion (Patel and Azzam, 2005; Cuzzocreo et al., 2009).

In total, forty participants were recruited, of whom two participants were considered drop-outs, while data from two participants were removed from the analysis to ensure that the groups were age-matched, leading to a total of thirty six participants ($N = 36$). In detail, the HC group consisted of 12 participants, the SCD group of 13 and the MCI group included 11 participants. The demographic characteristics of the participants can be found in Table 1. All groups exhibited a similar range of age and education. Kruskal-Wallis test revealed no group differences with regards to age and years of education (Table 1).

As the study was conducted during the pandemic (2021), solely fully vaccinated (validated vaccination certificates with verified app) participants were recruited. Moreover, after each participant's visit decontamination by experts took place at the SH to ensure the safety of all people involved.

2.1.2 Study design

Participants had the option of staying overnight at the SH or only for a daily visit. The study protocol consisted of five Tasks, of which three Tasks comprised ADL activities, namely, Task 1- Meal Preparation, Task 2 - Beverage Preparation and Task 3 - Snack Preparation (Figure 2). Two tasks consisting of meditation sessions were also included in the protocol (Task 4 - Mindfulness Based Stress Reduction –MBSR; Crane et al., 2017; Creswell et al., 2019), and Task-5 Kirtan Kriya meditation (Khalsa, 2015), where participants' performance during meditation was monitored using a portable Muse EEG device. The protocol and the study outline have been presented in Stavropoulos et al. (2021a) and

4 http://www.alzheimer-hellas.gr/index.php/el/

TABLE 1 Demographic characteristics of the participants ($N = 36$).

	HC	SCD	MCI	*p*
	$N = 12$	$N = 13$	$N = 11$	
Demographic characteristics				
Age in years	63.9 (6.4)	64.4 (6.4)	69.7 (6.4)	0.109
Gender (F:M)	11:1	9:4	8:3	
Years of education	13.8 (2.6)	14.6 (2.1)	12.9 (2.7)	0.292
Neuropsychological tests				
Mini Mental State Examination (MMSE) (Folstein et al., 1975)	29.25	27.85	26.00	**<0.001**
Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005)	26.83	25.54	20.64	**<0.001**
Functional rating scale for symptoms of dementia (FRSSD) Total Score (Hutton et al., 1998)	2.25	2.62	3.27	0.181
Functional and Cognitive Assessment Test (FUCAS) Total Score (Kounti et al., 2006)	42.00	42.00	44.36	**<0.001**
Rey-Osterrieth Complex Figure Test (ROCFT) Copy (Osterrieth, 1944)	35.25	33.00	30.23	**<0.001**
Rey-Osterrieth Complex Figure (ROCFT) Delayed Recall (Osterrieth, 1944)	18.50	20.19	10.86	**0.002**
Rivermead Behavioral Memory Test (RBMT) Immediate Recall (Wilson et al., 1989)	15.42	13.85	10.45	**0.003**
Rivermead Behavioral Memory Test (RBMT) Delayed Recall (Wilson et al., 1989)	13.83	11.96	7.55	**0.002**
Rey Auditory Verbal Learning Test (RAVLT) Total Score (Rey, 1964)	45.17	39.15	34.00	**0.025**
Trail Making Test (TMT) Part B (Tombaugh, 2004)	146.67	151.38	217.82	**0.045**
Verbal Fluency Test (FAS) (Kosmidis et al., 2004)	11.44	10.13	9.43	**0.009**
Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog) (Rosen et al., 1984)	9.47	11.96	16.58	**0.001**

Bold values denote statistical significance at the $p < 0.05$ or $p < 0.001$ level.

Lazarou et al. (2022) while the results of the meditation sessions have been reported in a separate publication (Ioulietta et al., 2023). The complete protocol and the full step by step description of each Task, as given to the participants, can be found in the Supplementary information. The total duration of the study (visit of first participant until visit of last participant) was approximately 3 months.

Upon arrival, participants were welcomed to the SH by the researchers and a detailed tour of the house followed. Afterwards, time for discussion and additional questions was planned and the study structure/protocol was again presented to the participants. Researchers then left the SH, and participants were encouraged to feel at home and perform the requested ADLs alone. For emergencies, they could contact the researchers via telephone or press one of the installed panic buttons. A psychologist- clinical research associate at CERTH was at all times available.

2.1.3 Participants' feedback (feasibility assessment)

At the end of the visit, questionnaires regarding study feasibility and technology evaluation were distributed to the participants, namely an overall study satisfaction questionnaire, the System Usability Scale (SUS), and the PANAS questionnaire assessing positive and negative affect (Supplementary information; Brooke, 1986; Watson et al., 1988).

2.2 Infrastructure

2.2.1 Smart home setting

The study was performed in the CERTH/ITI nZEB SH (Figure 3), a fully equipped, real domestic building, where participants can engage in real-world living scenarios and explore a plethora of innovative, smart IoT-based technologies. The SH can be used to test, validate and evaluate novel technologies from various fields, including but not limited to, Health, Energy, Big Data, Robotics and Artificial Intelligence (AI).

In this study, the SH environment was used to resemble and simulate the participants' home, with the installation of a number of sensors in every room allowing for unobtrusive monitoring of participants' ADLs. The available to the participants' spaces in the SH included one living room, a kitchen, a bedroom and two bathrooms.

2.2.2 IoT devices infrastructure

2.2.2.1 Installed sensors

IoT device selection resulted from extensive literature research and discussions with the partners of the RADAR-AD Consortium[5] (Owens et al., 2020; Stavropoulos et al., 2020). Also, focus groups with EWGPWD[6] and Alzheimer Europe[7] were assembled in order to rate the devices based on their features and potential usage and finalize the selection process (Stavropoulos et al., 2021b). Furthermore, an online meeting was organized (11/03/2021) to collect the EWGPWD's feedback on the fixed in-home sensors used in this study.

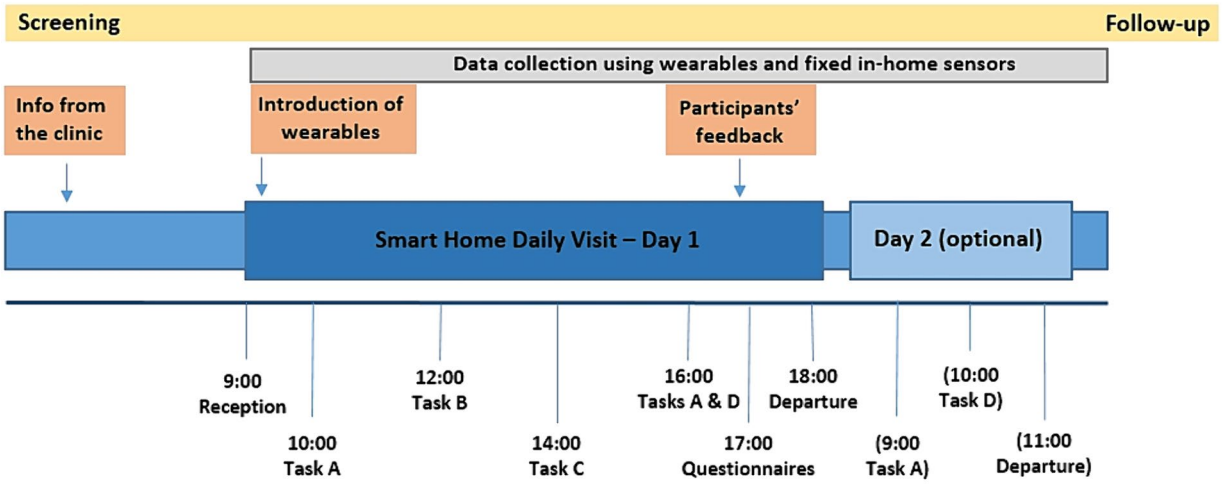
For the present study, commercially available Motion Sensors (quantity, $n = 8$) were added in every room of the SH to detect human presence. Furthermore, Door/Cabinet Sensors ($n = 8$) were placed on the main doors, as well as on the kitchen cabinets and drawers to

5 https://www.radar-ad.org/sites/radarad/files/2021-02/RADAR-AD%20device%20selection%20report.pdf

6 https://www.alzheimer-europe.org/about-us/european-working-group-people-dementia

7 https://www.alzheimer-europe.org/

A



B

	ADL Tasks	Step-by-step Task description	Time
	Task 1 - Meal Preparation	1. Open the cupboard labelled "Food" 2. Turn on the hot plate 3. ... 4. ... 5. Enjoy your meal!	13.00-15.00 **~30 min**
	Task 2 - Beverage Preparation	1. Open the cupboard with the label "Coffee-Tea" 2. Turn on the boiler machine 3. .... 4. Turn off the kettle 5. ...	10.00-10.30 and/or 16.00-16.30 **~10 min**
	Task 3 – Snack Preparation	1. Open the fridge 2. Open the cupboard labelled "Dishes" 3. Turn on the toaster 4. ... 5. ... 6. Wash the dishes	10.00-11.00 **~ 10 min** & 16.30-17.00 **~ 15 min**

FIGURE 2
(A) Overview of the complete study protocol and the structure of the daily visit with the optional overnight stay. (Task A – Beverage Preparation, Task B – Physical Exercise and Meditation Sessions, Task C – Meal Preparation, Task D – Snack Preparation). **(B)** Overview of the activities protocol, describing step-by-step the Tasks to be performed during the participants' visit (the full step-by-step description can be found in the Supplementary information).

signal if they were being opened and closed. Wall Plugs ($n = 6$) to measure consumption were added to small electrical appliances (e.g., kettle, toaster, hot plate) and four panic buttons were placed in the SH for emergencies. Examples of the installed fixed in-home sensors can be seen in Figure 3.

2.2.2.2 Raw data

The sensors generate data that consist of two types of time series, Signals, and Consumptions, both of which express the change of a device's status or metrics, respectively. Motion, Door/Cabinet sensors and Panic Buttons compose the Event time series, expressing with Boolean values the sensor's status (1 for Activated and 0 for Idle). The sensors are activated when a person interacts with them or with the environment (e.g., Entering a room activates the Motion Sensor, Opening the Cupboard activates the Cupboard's sensor, pressing the Panic Button sends the corresponding signal). Wall Plugs on the other hand, express the change of a home appliance's power consumption.

2.2.2.3 Hubs

The time series are generated via a small gateway device[8] designed to manage an entire SH system. Signal time series comprise signals from all sensors except for Wall Plugs, for which a Consumption time series is generated separately. In addition, the gateway device provides a REST API to serve the data to other services, such as the CARL Platform developed by our research team (Mpaltadoros et al., 2021).

2.2.3 Data collection and visualization

2.2.3.1 Data model

The CARL Platform (Care Ally: Data Collection and Analysis Platform for Assisted Living) is an end-to-end data collection and

8 https://www.fibaro.com/en/products/home-center-lite/

FIGURE 3
Top: The smart home site. Bottom: Examples from the sensor installation in the smart home (left: marked Wall Plugs, Door/Cabinet and motion sensors in the kitchen, right: Motion sensor in the bedroom).

analysis platform that allows integration with a continuously expandable list of commercially available wearable and IoT sensors and apps. Additionally, the platform offers a Visualization Dashboard for clinicians (real time data representation), to enable operational and clinical oversight across the entire lifespan of a study, in this way facilitating informed decision-making.

Integration of the gateway with the CARL Platform was achieved with the development of two components, the CARL RPi Client and an Adapter. The CARL RPi Client is a client service designed to detect the gateway on a local network and consume the generated time series in order to upload them to the CARL Platform. The Adapter was responsible for the authentication of the incoming data from CARL RPi Client instances and the serialization of the raw data. In this way, all data was transferred to the CARL Platform central database.

2.2.3.2 Visualization services

Once the raw data was saved in CARL Platform's database, it was processed to produce Event Objects, representing the duration of various events that occurred during the participant's visit (e.g., Cupboard Opened, Kitchen Presence, Hot Plate On). A clinician could then visualize these Event Objects through the dashboard, gaining an overview of all the participant's interactions with the environment (Figure 4).

2.3 Monitoring and synthesis of ADLs

2.3.1 From raw data to ADLs

In this section, a detailed description of the process followed to structure and transform the raw data into Tasks and ADLs is presented. An overview is given in Figure 5, where it can be seen how the raw signal and consumption data are converted to Events, while sequences of these Events are utilized to form ADLs.

2.3.1.1 Raw data to events

The production of Event objects based on Signal data is achieved by pairing the consecutive alterations of the device's status. When a Signal has a "newValue" of 1, a new Event object is generated, having as starting point the Signal's timestamp. The next Signal with a "newValue" of 0, will act as the ending point of the Event. An overview can be found in Table 2.

2.3.1.2 Raw consumption data to event

For Consumption based Events, we took into consideration that all devices, even when idle, still consume electrical power. Therefore, depending on the home appliance, we applied an empirical threshold, used to define when the home appliance was turned on and off. If the consumption value exceeded the set threshold, then an Event object was generated with the start time equivalent to the Consumption's timestamp. The next Consumption's timestamp with a value below the threshold, was used to mark the Event Object's end time.

2.3.1.3 Post-processing deviations

Due to the nature of the Cupboard/Door sensors, deviations were noticed in the duration of some related events (e.g., the cupboard did not fully close due to the brakes and extreme values were captured). On such occasions, the events were post-processed by inspecting each participant's Event objects from the sensor, to determine the distribution

10.3389/fnagi.2024.1375131

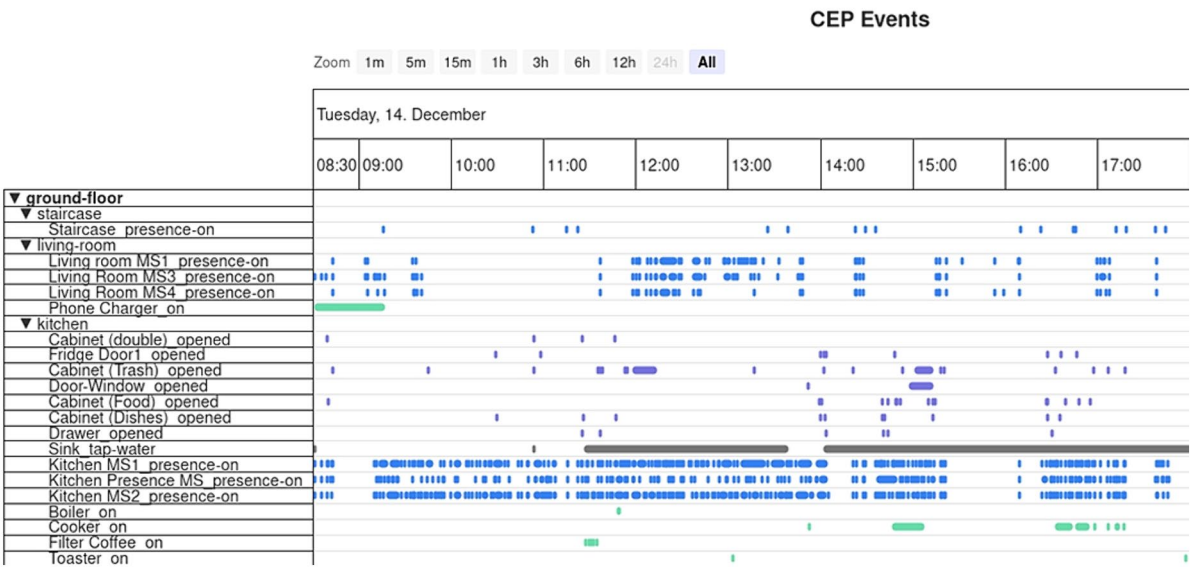


FIGURE 4
Event objects visualization using the CARL platform visualization dashboard.

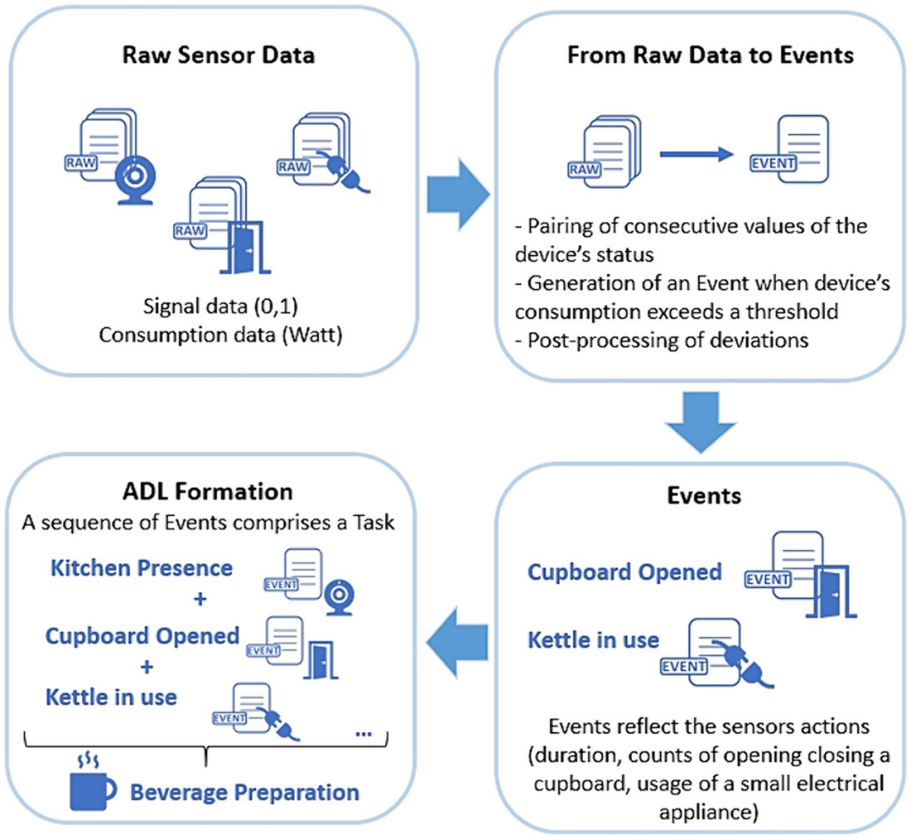


FIGURE 5
An overview of the transformation of raw sensor data into ADLs.

of the values. The duration values at the 75th percentile were then compared to the sum of the 75th percentile and Standard Deviation values. If the duration value exceeded the sum, we updated the end time of the Event to match the 75th percentile value. This process was applied until all duration values were lower than the sum value.

TABLE 2 Signal information obtained by the sensors (CARL: Care Ally: data collection and analysis platform for assisted living).

Signal information	Explanation
Id	Unique identifier
Timestamp	When it occurred in unix epoch
deviceID	Device's unique identifier
deviceType	Device's type
oldValue	Previous status
newValue	Current status

TABLE 3 Example of an ADL and the respective sensor activation sequence.

Example – Step by step description for Task 2 - beverage preparation	Sensors activated
• In the kitchen, fill the kettle with water and turn it on from the button • Open the cabinet labelled "Coffee - Tea" • … • Make sure you close the button from the kettle • After finishing drinking your coffee, wash the cup and the kettle with the dish sponge and leave them in the sink to dry	• Motion Sensor "Kitchen Presence" (ON) • Wall plug sensor "Kettle" (ON) • Door/Cabinet sensor "Coffee/Tea Cabinet" ON for a second then OFF • … • Wall Plug sensor "Kettle" (OFF) • Motion Sensor "Kitchen Presence" (OFF)

2.3.1.4 Events to ADLs

Through the CARL platform, it is possible to check whether the sensors were successfully activated by the participants compared to the task descriptions provided. An example of a step by step description and the respective expected sensor activations are presented in Table 3 for Task - 2 Beverage Preparation.

In order to form each of the three ADL Tasks, the use of a small electric appliance, depending on the task was considered necessary. In detail, for Task 1 – Meal Preparation the hot plate should be used, in Task 2 – Beverage Preparation the kettle was needed, while in Task 3 – Snack Preparation, the toaster was considered essential.

In Figure 6 the rationale of forming an ADL (Task 1 – Meal Preparation) is given. Initially, a home appliance based event ("Hot Plate Event" green bar, Figure 6) was detected. In order to take into account the event, its duration had to exceed a specific value. This was set by the researchers during the testing phase and served as a checkpoint (minimum duration for the hot plate $t = 10$ min, kettle $t = 2$ min and toaster $t = 5$ min). From there, thresholds were applied before and after the appliance's related event ("Threshold prior to Hot Plate event" and "Threshold after Hot Plate Event" Figure 6). The thresholds were determined after manual inspection of the data of all participants and were set for the hot plate at $t = 15$ min, the kettle at $t = 5$ min and the toaster at $t = 5$ min. All relevant Event objects occurring in between these thresholds (purple bars and lines, green bar, Figure 6) were clustered into one entity leading to an ADL (orange bar, Figure 6).

Apart from allowing the formation of the ADLs, visualization through the CARL platform offers a plethora of information on the performed tasks. For example, for Task 1 – Meal Preparation (Figure 6) it can be seen in which order the different sensors were activated and for how long during a specific point in time, while the participant performed the task. In detail, the orange bar shows the duration of the complete Task, information derived by considering all the individual sensors involved in the ADL performance as described above. The green bar shows the duration of the electrical appliance in use, in this case the hot plate, derived by the consumption observed

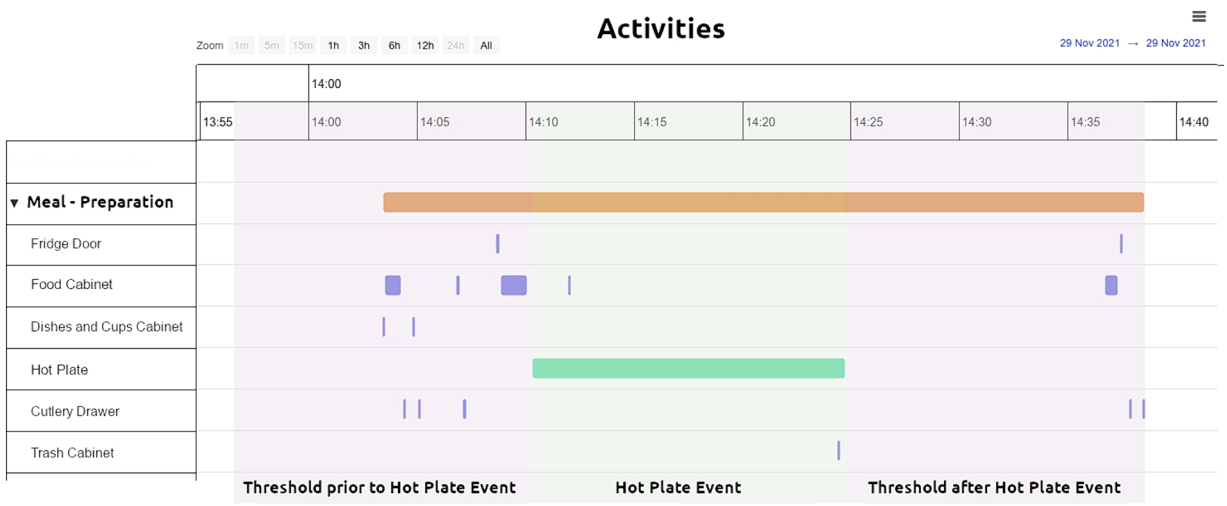


FIGURE 6
ADLs based on the collected data - threshold applied to small electrical appliances consumption (e.g., hot plate). Task – 1 Meal Preparation activity visualized through the CARL platform.

TABLE 4 Feature description and the naming convention followed for the sensors used in each Task/ADL.

Feature	Description		Data type
Activity_name	ADL's Name (i.e., Task 1 – Meal Preparation, Task 2 – Beverage Preparation, Task 3 – Snack Preparation)		Text
Activity_duration	The time needed (duration) to perform an ADL (Task 1, Task 2, Task 3). All sensors comprising the ADL are taken into account [The time between the start_time of the first Event Object and the end_time of the last Event Object]		Seconds
Number_of_steps	The total number of sensors activated during the performance of an ADL		Integer
Count_<name_of_sensor>	The number of times, number of repetitions (count) a Door/Cabinet sensor was activated		Integer
Sum_<name_of_sensor>	The time (SUM duration) of a sensor being activated		Seconds
Avg_<name_of_sensor>	The time (AVG duration) of a sensor being activated [SUM duration divided by the number of repetitions (sum duration and average duration are identical if sensor was used/activated once)] [sum_<name_of_sensor>/count_<name_of_sensor>]		Seconds
Sum_<inaction_time>	Time period during a Task where the participant did not activate any sensors [activity_duration – SUM (sum_<sensors>)]		Seconds
Sensor name	**Type of sensor**	**Sensor description**	**Task**
Coffee – Tea cabinet	Door/cabinet sensor	Door/drawer/cabinet opening - closing	2
Dishes and cups cabinet			1, 2, 3
Cutlery drawer			1, 2, 3
Food cabinet			1, 3
Trash cabinet			1, 2, 3
Fridge door			1, 3
Hot plate	Wall plug	Consumption monitoring	1
Kettle			2
Toaster			3
Kitchen, living room, bedroom, bathrooms, hallways	Motion sensors	Presence/motion capture	ADLs were performed in the Kitchen

during this period of time. The more frequent (due to number of repetitions) and thinner (due to shorter duration) purple signals, show the various cabinets and drawers opened and closed during the execution of the Task.

2.3.2 ADL features

Our intention with analysing the raw sensor data into events and ADLs associated with specific Tasks, was to enable the extraction of representative features characterising an ADL, and use these features to detect differences between the groups of participants (HC, SCD and MCI).

More specifically, we can see the time of the day the participants performed each task, the duration of each activity (in seconds), as well as the duration each appliance was in use (in seconds), or the time a cabinet was left open, and the number of times a sensor was activated (number of repetitions, count for, e.g., opening a cabinet). Furthermore, apart from these primary derived features, a secondary feature/by-product was investigated, namely the "Inaction Time" which refers to the time recorded between sensor activation. For this, the durations between different sensor signals were added up and subtracted from the total activity duration creating the feature "Inaction Time."

In Table 4 the description of each feature is given, along with the naming convention followed for each sensor, the sensor type and in which of the Tasks they were utilized. All this information was exported in the form of a csv file, to facilitate statistical analysis.

2.3.3 Validation

To ensure the sensors' and platform's effectiveness and reliability, during the study, information was gathered from the participants by the researchers in the form of free text notes, regarding the performed ADLs (completed Tasks, approximate time of the day performed) and used as ground truth. A comparison between the ground truth and the activities identified by the platform was made. Differences in the number of activities recorded by the platform and the available ground truth data could be attributed to power and internet outage or sensor connectivity issues. In detail, one "Meal Preparation" and one "Beverage Preparation" tasks were missing from the platform due to unexpected power outage in the SH. For two "Snack Preparation" tasks (performed the same day), the platform collected only data from the Wall Plug sensor, while the Door/Cabinet sensors were unresponsive.

2.3.4 Statistical analysis

With the dataset containing all information on the various features per task at hand, we proceeded to compare the performance on each ADL, among the three groups at the level of significance $p = 0.05$. Descriptive analysis and statistical

analysis were performed using SPSS v25.0 for Windows (IBM Corporation, Armonk, NY, United States). Descriptive analysis was performed to depict participants' data, while statistical analysis was carried out to locate differences in the various activities and the individual features.

For assessing the normality assumption for continuous variables we used the Kolmogorov–Smirnov test. As the depended variables were not normally distributed, and due to the small sample size available, non-parametric tests were selected (Mishra et al., 2019). Between groups comparisons were made using the Kruskal-Wallis test. For examining the potential statistical significance between two independent groups (e.g., HC versus SCD), the Mann–Whitney test was used. Furthermore, the Area under the Curve (AUC) was also examined.

3 Results

3.1 Exploring ADLs – task comparison between groups

The assumptions formed in the present study were shaped around the expectation that more cognitively impaired participants will exhibit different behavioural patterns compared to healthy controls. These differences can be attributed to functional deterioration, as AD is characterized by the impairment of cognitive functions and increasingly poorer ADL performance.

Specifically, it is expected that the differences in ADL performance will be observed in the overall time needed to complete an ADL, in additional steps made and repeated actions noted (e.g., opening/closing a cabinet more frequently).

Consequently, the features considered meaningful to explore these assumptions, as derived from feature engineering of the collected sensor data, include number of steps to complete an ADL, activity duration, sensor activation duration, number of sensor activations, and inaction time.

Descriptive statistics and results for the statistical tests are given for all features in the Supplementary information. It is noted that while all results are commented in the text, only the more prominent for discussion features are presented in Figures and Tables to provide the reader with a clearer overview.

3.1.1 Task completion

Participants were asked to complete three Tasks as entailed in the protocol. Three tasks were completed by 33% of the participants of the HC group (4/12), 23% of the SCD (3/13) and 18% (2/11) of the MCI group. Two tasks were performed by the majority of the SCD group (61.5%, 8/13), approximately half the participants of the MCI group (55%, 6/11) and by 42% (5/12) of the HC group. Furthermore, 27% (3/11) of the MCI group completed only one task, whereas the percentages are 25% (3/12) and 15% (2/13) for the HC and the SCD group, respectively.

In detail, it is noted that 11/12 HC (91.67%), 11/13 SCD (92.31%) and 7/11 MCI (63.64%) performed the activity "Meal Preparation." The activity "Beverage Preparation" was performed by 10/12 HC (76.92%), 11/13 SCD (92.85%) and 6/11 of the MCI (61.53%). Only twelve participants performed the Task "Snack Preparation," in detail, 5/12 HC, 4/13 SCD and 3/11 MCI.

3.1.2 Number of steps

The estimated number of steps needed to complete Task 1 – Meal Preparation, according to the step-by-step task description is ten. The mean number of steps for each group was found to be 14.7 (SD = 3.8) for HC, 14.6 (SD = 5.6) for SCD and 18.8 (SD = 8.2) for the MCI group, showing no differences between the HC and SCD groups, and a larger number of steps for the MCI group. No statistical significance was noted (Kruskal-Wallis test $p = 0.437$).

For Task 2 - Beverage Preparation, the description included six sensor activation steps, while participants performed HC = 7 (SD = 1.5), SCD = 7.1 (SD = 1.3), MCI = 7.6 (SD = 3.9) steps.

The three groups needed approximately the same number of mean steps to complete Task 3 – Snack Preparation [HC = 9.8 (SD = 2.9), SCD = 11.3 (SD = 4.9), MCI = 10.0 (SD = 3.6)]. It is noted that the protocol lists six sensor activation steps for this Task.

3.1.3 Activity duration

Furthermore, the time needed to complete a Task was assessed. The distribution of the collected data is presented in Figure 7A for the three Tasks. The results of the Kruskal-Wallis test for the "Activity Duration" feature for the three Tasks showed a statistically significant difference across the three groups of participants at a $p = 0.05$ level, in Task 1 – Meal Preparation [H (2) = 7.607, $p = 0.022$] (Table 5). No statistically significant difference was noted for "Activity Duration" in Tasks 2 and 3.

Afterwards, in order to determine the groups between which discrimination was possible in Task 1 – Meal Preparation, Mann Whitney test was performed, showing that the duration was statistically significant longer for the SCD group compared to HC (U = 29.00, $p = 0.040$), and also for the MCI group compared to HC (U = 9.00, $p = 0.015$). No differentiation was possible between the SCD and the MCI group for Task 1.

3.1.4 Individual sensors

As each Task consists of a synthesis/composition of Events, signalled by different sensors, it was considered important to investigate next the activation duration as well as the number of activations marked for the individual sensors. It is noted that six sensors were placed to monitor Task 1 – Meal Preparation, five for Task 2 – Beverage Preparation, and six for Task 3 – Snack Preparation (Table 4 in previous section).

3.1.4.1 Sensor activation duration

Kruskal Wallis test was performed for all available sensors regarding the features "sum_<name_of_sensor>" and "avg_<name_of sensor> (Supplementary information).

A statistically significant difference was found only for the duration of the sensors placed on the Fridge Door and the Food Cabinet in Task 3 – Snack Preparation, while a weak trend was observed for the Fridge Door sensor in Task 1 – Meal Preparation (Table 5). Mann Whitney tests for the sensors of Task 3, revealed a trend between the HC and MCI groups, showing that the MCI group noted longer durations when utilizing the Fridge Door and the Food Cabinet during the Snack Preparation task. The data distribution of the abovementioned sensors can be found in Figure 7B.

Regarding the use of the small electrical appliances, no differentiation was possible. Boxplots showing the distribution among

10.3389/fnagi.2024.1375131

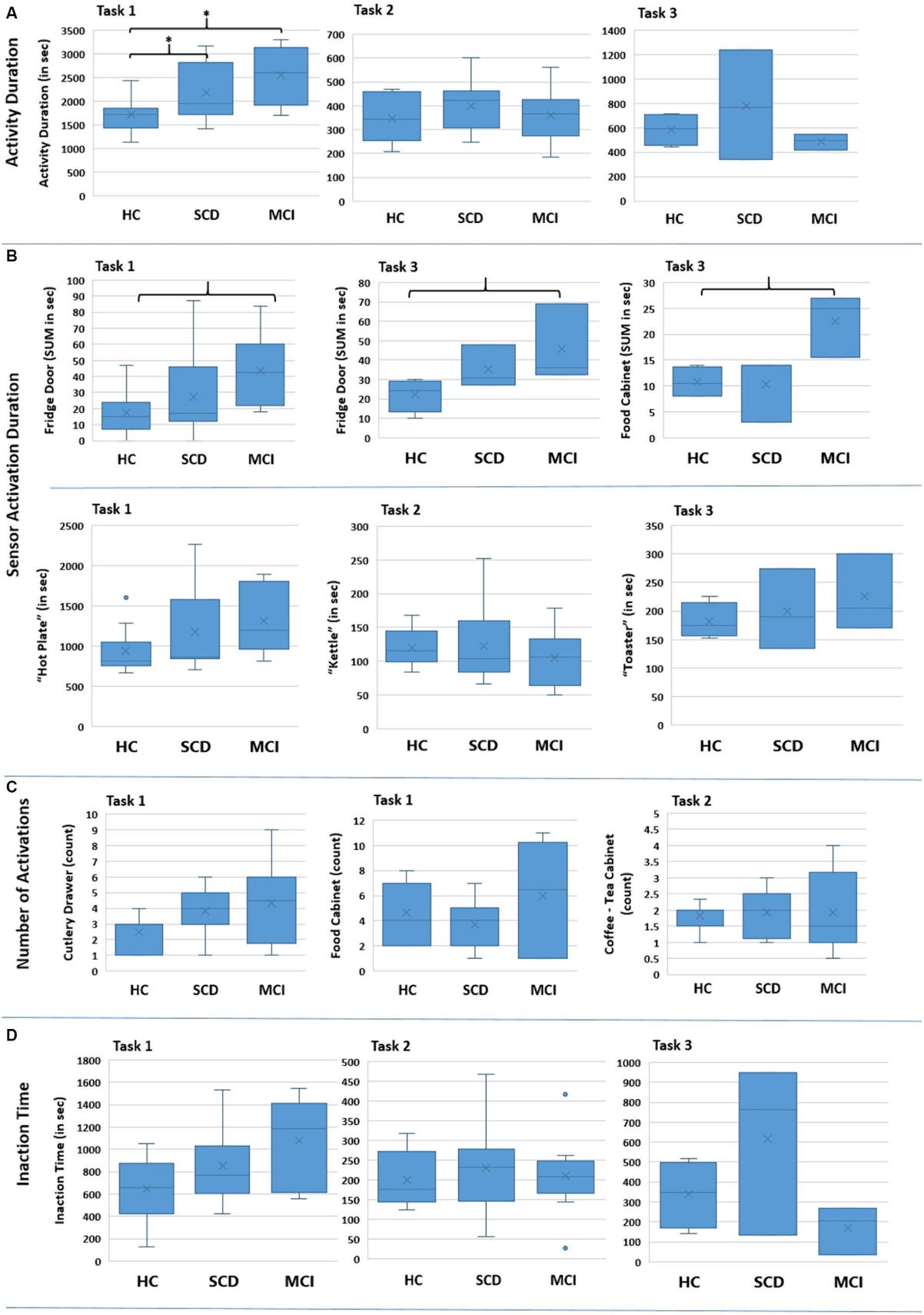


FIGURE 7
Boxplots showing the distribution of the collected data for the various features for the three groups regarding each Task (Task 1 – Meal Preparation, Task 2 – Beverage Preparation, Task 3 – Snack Preparation). The * in **(A)** indicates the group pairs where a statistical significance at p=0.05 level was found. The brackets in **(B)** indicate the group pairs where a weak trend (p=0.057) was found. No statistical significance was found in **(C)** and **(D)** between groups.

TABLE 5 Descriptive statistics (mean value and standard deviation, given in seconds and count according to each feature), Kruskal–Wallis and Mann–Whitney *p* values for the explored features regarding the performed Tasks (Task 1 – Meal Preparation, Task 2 – Beverage Preparation, Task 3 – Snack Preparation).

Task	Sensor	HC	SCD	MCI	Kruskal Wallis *p*-value	Mann–Whitney U-test *p*-value		
		Mean value (standard deviation)				HC versus SCD	HC versus MCI	SCD versus MCI
Feature "Activity Duration" (in seconds)								
1	All sensors comprising the corresponding ADL are taken into account	1710 (349)	2,180 (604)	2,546 (619)	**0.022**	**0.040**	**0.015**	0.350
2		346 100	400 (104)	362 (106)	0.566	–	–	–
3		586 (135)	780 (450)	487 (66)	0.546	–	–	–
Feature "Sensor Activation Duration" (in seconds)								
1	Fridge door	17 (14)	27 (26)	44 (24)	0.074	–	–	–
3	Fridge door	22 (9)	35 (11)	46 (20)	**0.046**	–	0.057	–
3	Food cabinet	11 (3)	9 (8)	23 (6)	**0.050**	–	0.057	–
1	Hot plate	941 (282)	1,176 (546)	1,314 (444)	0.139	–	–	–
2	Kettle	120 (29)	123 (53)	117 (56)	0.963	–	–	–
3	Toaster	182 (31)	200 (70)	226 (67)	0.554	–	–	–
Feature "Number of Activations" (count)								
1	Cutlery drawer	2.45 (1.13)	3.82 (1.60)	4.33 (2.80)	0.093	–	–	–
1	Food cabinet	4.64 (2.46)	3.73 (1.95)	6.00 (4.38)	0.517	–	–	–
2	Coffee – Tea cabinet	1.81 (0.13)	1.92 (0.21)	2.03 (0.38)	0.984	–	–	–
Feature "Inaction Time" (in seconds)								
1	Time period during a Task where no sensors were activated	647 (284)	854 (305)	1,081 (401)	0.148	–	–	–
2		200 (70)	230 (111)	210 (103)	0.762	–	–	–
3		339 (171)	616 (428)	169 (120)	0.394	–	–	–

Bold values denote statistical significance at the $p < 0.05$ level.

groups are presented in Figure 7B, while descriptive statistics and the results from the statistical tests are also included in Table 5.

3.1.4.2 Number of activations (count per sensor)

No interesting finding could be noted here. Even though this feature could be connected with the performed number of steps, where for Task 1 – Meal Preparation the MCI exhibited a larger number of mean steps, no statistically significant difference could be found. Indicative examples of the data obtained can be seen in Figure 7C and Table 5.

3.1.5 Inaction time

The composite feature "Inaction Time," aiming to capture the time participants spent during an activity without activating a sensor (e.g., due to wandering, considering their next action), did not yield any differences between groups. Considerable overlapping between groups is noted for "Inaction Time," and no differences emerged from the Kruskal-Wallis tests performed (Figure 7D; Table 5).

3.2 Sensitivity and specificity

The potential utility of the three ADL tasks as objective markers to distinguish an individual's cognitive condition (SCD, MCI) compared to HC by testing Sensitivity and Specificity among the groups (Table 6) was investigated.

In general, an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without the disease or condition based on the proposed test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding (Hosmer and Lemeshow, 2000; Mandrekar, 2010).

In detail, we managed to successfully discriminate HC from SCD in Task 1 – Meal Preparation (AUC = 76%, Sensitivity = 0.82 and Specificity = 0.64) regarding the "Activity Duration" feature. Also, we managed to discriminate HC from MCI in Task 1 – Meal Preparation (AUC = 86%, Sensitivity = 0.83 and Specificity = 0.82) and Task 3 – Snack Preparation (AUC = 75%, Sensitivity = 0.75 and Specificity = 0.67). Interestingly, no discrimination could be made between the SCD and MCI groups. The "Activity Duration" feature of the Meal Preparation Task can distinguish between HC-MCI and HC-SCD with acceptable robustness.

3.3 Usability and satisfaction questionnaires

The overall experience as perceived by the participants during the study in the SH was assessed with a study satisfaction questionnaire that referred to the visit, the tasks, the time needed to complete the tasks and the level of difficulty (Supplementary information). No difference between the three groups could be noted regarding the given feedback. The majority of the participants (72%) when asked if they were satisfied with their participation in the study, replied with

TABLE 6 Sensitivity and specificity of the three ADL tasks (Task 1, Task 2 and Task 3) regarding the feature "Activity Duration" for discriminating between groups.

Feature "activity duration"		AUC	Threshold value (in seconds)	Sensitivity	Specificity
HC versus SCD	Task 1 – Meal preparation	**0.76**	1715	0.82	0.64
	Task 2 – Beverage preparation	0.62	342	0.67	0.56
	Task 3 – Snack preparation	0.67	700	0.67	0.75
HC versus MCI	Task 1 – Meal preparation	**0.86**	1924	0.83	0.82
	Task 2 – Beverage preparation	0.58	352	0.70	0.56
	Task 3 – Snack preparation	**0.75**	500	0.75	0.67
SCD versus MCI	Task 1 – Meal preparation	0.65	1977	0.83	0.55
	Task 2 – Beverage preparation	0.59	403	0.58	0.80
	Task 3 – Snack preparation	0.67	522	0.67	0.67

The sensitivity and specificity scores corresponding to the cut-off thresholds alongside with the AUC. Bold values suggest acceptable discrimination (0.7–0.8) or excellent discrimination (0.8–0.9), (Hosmer and Lemeshow, 2000; Mandrekar, 2010).

"Extremely satisfied." The participants perceived the study as "Extremely appealing" (60%), "Very appealing" (22%) and "Appealing" (14%). The time planned for the Tasks was found to be sufficient with participants commenting that they did not need more time to complete the activities (94%). The instructions and task descriptions were unanimously found to be extremely easy to read and understand, and the labels placed to mark the different cabinets (labelled "Food," "Cutlery") were perceived as very useful. No technical issues and no issues of any other nature were noted during the participants' visit (e.g., problems with the sensors, person wanting to terminate participation). All participants replied with "No" when asked if any skills were required to interact with the proposed technologies. Describing the overall visit, all participants replied that their participation was a positive experience and no feeling of inconvenience was noted (e.g., stress, depression, anxiety). Additionally, the mean scores (M) per group, for the SUS questionnaire (scores can range from 0 to 100) (Brooke, 1986), revealed excellent overall usability [HC: M = 94 (SD = 5.8), SCD: M = 92.9 (SD = 4.7), MCI: M = 93.9 (SD = 5.2)]. The PANAS questionnaire, designed to measure emotional experience (namely positive affect, PA and negative affect, NA was utilized; Watson et al., 1988). Respondents are asked to indicate the extent to which they have experienced each emotion (e.g., excitement, sadness) over a specific period of time rating them on a scale from 1 to 5 (PA and NA can range from 10 to 50). The participants showed acceptable positive and negative affects, while no differences between groups were observed [HC: $M_{PA} = 36$ (5), $M_{NA} = 20$ (5), SCD: $M_{PA} = 35$ (3), $M_{NA} = 19$ (6), MCI: $M_{PA} = 36$ (5), $M_{NA} = 21$ (6)].

4 Discussion

The Smart Home, simulating a domestic residence, offers a unique environment allowing for controllable experimental conditions. Through data collection using non-intrusive fixed in-home sensors in the

CERTH-ITI SH, and instructing participants to follow a protocol listing a number of ADLs, we aimed to capture, quantify and assess ADL performances, as these can lead to insightful measures for functional deterioration. Through this ecologically valid assessment, we aimed to detect changes between three different cohorts, namely HC, participants with SCD and participants with MCI. Visualization of the collected data and extraction of meaningful features in the form of a dataset available for analysis was possible by utilizing the CARL platform.

This preliminary investigation demonstrated that SH technologies present an opportunity for an unbiased and real-world evaluation of ADLs in individuals with SCD and MCI. The study allowed for the assessment of not only whether a task is accomplished but also how it is carried out.

Discussing protocol adherence and number of overall completed ADLs, it appeared that participants did not follow precisely the provided protocol with the step-by-step task descriptions, but proceeded with the ADLs in a more intuitive way. Additionally, commenting on the number of steps needed to complete a Task, only for Task 1 – Meal Preparation was a small difference observed in the mean number of steps for the MCI group compared to the HC and SCD groups. In Tasks 2 – beverage preparation and 3 – snack preparation the three groups performed similar number of steps.

The correctness of the executed steps may not be easily assessed, using simple statistical analysis methods, as the step sequence differs not only between groups but also notably, within groups as well. However, as participants proceeded with the Task execution in a freely manner, the observations made are in the context of real-environment monitoring and allow real-life evaluations. Additionally, as commented in Jekel et al. (2016), we should consider that there could be significant individual variability in performing a task in a correct manner, for this, it can be overall argued, if correctness of steps can pose a useful feature. Also, in the work of Lundström et al. (2016) guidelines provided to participants for performing tasks (e.g., prepare breakfast, get hot drink, prepare dinner), were written in a simplified manner to allow for natural variation.

It is noted that overall the HC and SCD groups performed, respectively, 70 and 74% of the expected tasks and the MCI group 58%. Specifically for Task 1 – Meal Preparation, the MCI group exhibited the lowest number of performance compared to the other groups (11/12 HC, 12/13 SCD, 7/11 MCI). No plausible justification could be derived for this discrepancy.

Regarding activity duration, Task 1 – Meal Preparation, yielded differences between the groups, which constitutes an interesting finding. It was considered that the more elaborate task of preparing a hot meal, due to its added complexity, was able to highlight the groups' differences attributed to functional decline due to cognitive

impairment. Specifically, comparison of participants' performances in Task 1 led to statistically significant differences between groups, namely between HC versus SCD, and HC versus MCI, based on the time needed to complete the task. It is noteworthy that no differentiation could be made between SCD and MCI participants.

The meal preparation task has been investigated also in a different context, in the work of Atkins et al. (2015, 2018), where the Virtual Reality Functional Capacity Assessment Tool (VRFCAT) was used. Discriminating healthy older adults from older adults with SCD was possible, as the latter noted a statistically significant larger amount of time to complete the given tasks and performed more errors.

Furthermore, ROC Curve values were encouraging for the Task 1- Meal Preparation, reaching 86% in the classification of HC vs. MCI and exciding 70% in the classification of HC vs. SCD. This is a promising finding, as available neuropsychological tests do not discriminate SCD from HC (Sikkes et al., 2009; Kaur et al., 2016).

On the other hand, the more straightforward / simple tasks of preparing a beverage (Task 2), and a snack (Task 3) were not able to show between groups differences. This is in accordance with existing literature. In Jekel et al. (2016), the coffee and sandwich preparation tasks were also not able to differentiate the HC and MCI participants, while in Karakostas et al. (2020) assessing various ADLs, no difference could be observed between HC and MCI for the tea preparation task. As has been commented in Jekel et al. (2016), these tasks could be considered as not highly cognitive demanding.

Additionally, regarding the individual sensors, only the ones placed on the Fridge Door and the Food Cabinet (both Fridge and Food Cabinet entailing a variety of different products) could show some difference between groups in their utilization (weak trends). Again, we are of the opinion that the fridge and the cabinet containing a number of products could be considered as the more complicated to handle.

The feature "Inaction Time" was considered promising as it was assumed that cognitive impairment and functional decline could lead to increased wandering time between actions due to possible disorientation (Coughlan et al., 2018). While the participants' data distribution for "Inaction Time" in Task 1 – Meal Preparation showed this expected tendency, no statistical significant difference was observed.

The duration of utilizing the small electrical appliances was compared between groups. Since the activation of these appliances was seen as a requirement for the formation of the ADLs (the ADLs were built around the data collected from the small electrical appliances), it was important to determine if this factor predominantly influenced the overall composite ADL duration feature. However, no statistically significant difference could be observed between groups.

In general, for many of the collected sensor data, descriptive statistics revealed an initial trend that MCI participants (and in some cases SCD participants) exhibit longer durations than HC, but significant overlapping exists between the groups not allowing further comparisons. Regarding the features addressing aspects besides duration, like the number of steps needed to complete a Task, the number of repetitions in utilizing, e.g., specific cupboards, could not be used to differentiate the groups. We are of the opinion that these features are reflecting actions not cognitive demanding and are not granular enough to highlight differences. For this, further feature exploration is needed to gain additional markers from the performed ADLs.

The present study shows that implementing new technologies that are able to detect subtle changes in cognitive and functional patterns may allow earlier diagnosis, even at the point when memory functions are still intact, such as the SCD stage.

While studies on activity recognition from collected sensor data are available in the literature (Bouchabou et al., 2021), there is limited research on efforts of quantifying and comparing the performed ADLs among early stages of cognitive impairment (Atkins et al., 2018), while only a comparison between a small number of HC and MCI participants has been attempted so far (Stucki et al., 2014; Jekel et al., 2016; Stavropoulos et al., 2016; Urwyler et al., 2017; Karakostas et al., 2020).

Moreover, there is scarce evidence for real-life, smart home-based use of technologies for early detection of dementia, and no approach is yet perceived as mature enough (Piau et al., 2019). An exception can be considered the Collaborative Aging Research Using Technology (CART) Initiative, a multi-site, nationwide project (Thomas et al., 2021; Bernstein et al., 2022). The study uses multiple embedded sensing technology and diverse data to support research in the field of health and independent living, focusing on older adults from various communities. However, as the authors note, further proof is needed on the precision, accuracy and reliability of these novel outcome measures before home-based sensor technologies can be included in clinical trials and utilized in the monitoring of chronic diseases (Thomas et al., 2021).

A frequent constraint in the majority of studies that evaluate SH technologies for monitoring ADLs, is their lack of focus on participants' acceptance of the devices, as indicated by a recent systematic review (Lawson et al., 2023). Along with the fact that elderly participants are not very keen on using smart technologies (Tiersen et al., 2021; Wei et al., 2023), participants views need to be considered when introducing new technologies. The present study and the proposed technologies were evaluated by the participants, and were regarded as feasible. Participants answered in a positive manner when asked a number of questions regarding their experience and their stay, the sensors and technologies utilized, while they did not experience any issues.

The study has some limitations that need to be acknowledged. While the sample size (N = 36) could be considered sufficient, considering the exploratory nature of the study and existing literature (Hayes et al., 2008; Petersen et al., 2015; Seelye et al., 2020) it is noted that, as some participants did not complete all tasks listed in the protocol, the dataset was further decreased. For example, a number of people, independently of their group, did not perform the Task 3-Snack Preparation activity (7/12 HC, 9/13 SCD and 8/11 MCI). This could be attributed to the fact that as participants stated, "They were not hungry," or "preferred to rest some more" and "explore the Smart Home's premises instead." This led to a restricted dataset available for Task 3 for analysis, the findings of which must be viewed with caution.

Additionally, it is noted that, as participants visited a new, unknown to them environment, this could also have affected the way they performed the various ADLs. Nevertheless, effort was made to simulate a real domestic environment while also adequate time was provided to the participants to feel comfortable in the house and discuss any concern with the researchers.

During feature extraction, conversion of raw data to events and activities involved refinement through post-processing. Even though all data processing was performed in a systematic manner and is described in the text, and a validation of the ADLs derived from the

sensors was performed using collected information as ground truth, in the interest of thoroughness this is acknowledged as a potential limitation.

Also, the use of flood sensors was investigated, installed in the kitchen (sink) and the bathroom (sink and flush). However, as the sensors are designed to detect water leaks and flooding, the necessary adjustments made to the sensors to monitor water usage instead, did not allow robust and continuous data collection. For this, the sensors were not included in the study.

Furthermore, it is noted that, as biofluid biomarkers were not collected for all participants, the etiology of the MCI and SCD group cannot be distinguished (amyloid positive vs. amyloid negative).

A limitation of the study, to be addressed in future work, constitutes the absence of a comparison/correlation to relevant conventional measures of function [e.g., the informant-based Amsterdam IADL questionnaire (Sikkes et al., 2012), the Naturalistic Action Test (Seligman et al., 2014)].

Finally, regarding the study's feasibility assessment, as researchers were present while participants filled out the questionnaires, possible bias could occur.

The herein presented SH study provides a proof-of-concept for the feasibility of identifying, quantifying and assessing ADLs and differentiating known-groups via monitoring their performance. It is evident that new tools will be required to assess and evaluate clinically significant changes (Atkins et al., 2015; Gold et al., 2018). The inclusion of people in preclinical stages of AD, constitutes an important step towards the advancement of digital biomarkers.

5 Conclusion

Participants spent a day in CERTH-ITI's Smart Home, a controlled environment that simulates a fully functional house, and were asked to perform a number of ADLs according to a given protocol. The results proved the differentiation among the HC group compared to the SCD and the MCI groups considering the feature "Activity Duration" in Task 1 - Meal Preparation. Task 1 can be considered more complex compared to Task 2 - Beverage Preparation and Task 3 – Snack Preparation.

The distinction of the SCD from the HC group, constitutes an important finding, as conventional assessments (neuropsychological questionnaires) note no difference between these groups. Furthermore, the differentiation of HC and MCI participants, as documented in the existing literature, confirms the study design and the methodology followed. Additionally, it is interesting to note that no significant group differences could be observed between the SCD and the MCI groups.

These findings further support the interest and need to include people in preclinical stages of dementia in current research. Furthermore, the study was proven feasible, with participants expressing positive feedback for the study and the technologies used.

Access to this information, paves the way for detection of behavioural patterns and deviations allowing for early observation of deterioration in function. This ecologically valid study provides evidence that ADL performance can be utilized and further evolved to develop clinically relevant digital biomarkers. These biomarkers could serve for monitoring participants in at-home settings, participant stratification as well as secondary endpoints in clinical trials to complement established outcome measures.

Starting from these encouraging findings, further research would be needed to determine the long-term reliability and predictive value of the proposed assessment tools in the clinical practice. Consecutive data collection on the executed ADLs over an extended period of time, would allow us to monitor behavioral patterns of the individuals in depth, identify personalized thresholds and highlight potential functional deterioration. Additionally, in this way, other factors could be controlled for and tested. For example, measures of sleep duration and quality could be incorporated (by using wearable devices, or pressure sensors placed underneath the mattress on the bed) to better understand their influence on ADL performance. A longitudinal study could evaluate and strengthen the presented findings and provide a useful tool, to serve as a secondary endpoint in drug trials on the therapeutic efficacy of prescribed drugs.

Furthermore, while study centers are not widely available for healthcare research, we envision that as technology continues to evolve and becomes increasingly part of our everyday life, the suggested assessment could be implemented in home environments, facilitating the inclusion of people in rural areas. In detail, the integration of smart devices and appliances, outfitted with microprocessors and WiFi access, is steadily gaining prominence within domestic settings. This reflects a significant shift towards the adoption of interconnected technologies in everyday life. The proposed approach is scalable and cost-effective. The protocol deploys commercially available sensors, indicating its practicality and accessibility. Additionally, the developed CARL platform is device agnostic, allowing the integration of different sensors and demonstrating flexibility in technological advancements.

As a first step towards the implementation and exploration of testing this protocol at a home environment, another sub-study realized in the context of the RADAR-AD project was set up to explore the feasibility of such an approach. The fixed in-home sensors were placed in participants' homes and data collection was ongoing for 4 weeks.

Data availability statement

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the Centre of Research and Technology Hellas (CERTH) and the Scientific and Ethics Committee of the Greek Association of Alzheimer's Disease and Related Disorders (GAADRD). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Group members of RADAR-AD

Dag Aarsland, Halil Agin, Vasilis Alepopoulos, Alankar Atreya, Sudipta Bhattacharya, Virginie Biou, Joris Borgdorff, Anna-Katharine Brem, Neva Coello, Pauline Conde, Nick Cummins, Jelena Curcic, Casper de Boer, Yoanna de Geus, Paul de Vries, Ana Diaz, Richard

Dobson, Aidan Doherty, Andre Durudas, Gul Erdemli, Amos Folarin, Suzanne Foy, Holger Froehlich, Jean Georges, Dianne Gove, Margarita Grammatikopoulou, Kristin Hannesdottir, Robbert Harms, Mohammad Hattab, Keyvan Hedayati, Chris Hinds, Adam Huffman, Dzmitry Kaliukhovich, Irene Kanter-Schlifke, Ioannis Kompatsiaris, Ivan Koychev, Rouba Kozak, Julia Kurps, Sajini Kuruppu, Claire Lancaster, Robert Latzman, Ioulietta Lazarou, Manuel Lentzen, Federica Lucivero, Florencia Lulita, Nivethika Mahasivam, Nikolay Manyakov, Emilio Merlo Pich, Peyman Mohtashami, Marijn Muurling, Vaibhav Narayan, Vera Nies, Spiros Nikolopoulos, Andrew Owens, Marjon Pasmooij, Dorota Religa, Gaetano Scebba, Emilia Schwertner, Rohini Sen, Niraj Shanbhag, Laura Smith, Meemansa Sood, Thanos Stavropoulos, Pieter Stolk, Ioannis Tarnanas, Srinivasan Vairavan, Nick van Damme, Natasja van Velthogen, Herman Verheij, Pieter Jelle Visser, Bert Wagner, Gayle Wittenberg, and Yuhao Wu.

Author contributions

MG: Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Validation, Visualization. IL: Writing – original draft, Writing – review & editing, Conceptualization, Formal analysis, Investigation. VA: Data curation, Investigation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Formal analysis. LM: Writing – review & editing, Investigation. VO: Writing – review & editing, Investigation. TS: Conceptualization, Funding acquisition, Supervision, Writing – review & editing, Methodology. SN: Conceptualization, Funding acquisition, Supervision, Writing – review & editing, Methodology, Project administration. IK: Conceptualization, Funding acquisition, Supervision, Writing – review & editing, Methodology, Project administration. MT: Investigation, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The RADAR-AD project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement no 806999. This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Programme and EFPIA and Software AG. This communication reflects the views of the RADAR-AD consortium and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Acknowledgments

We thank all past and present RADAR-AD consortium members for their contribution to the project.

Conflict of interest

The 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2024.1375131/full#supplementary-material

References

Amiribesheli, M., and Bouchachia, H. (2018). A tailored smart home for dementia care. *J Ambient Intell Human Comput* 9, 1755–1782. doi: 10.1007/s12652-017-0645-7

Anna-Katharine, B., Sajini, K., Casper, d. B., Marijn, M., Ana, D.-P., Dianne, G., et al. (2023). Digital endpoints in clinical trials of Alzheimer's disease and other neurodegenerative diseases: challenges and opportunities. *Front. Neurol.* 14:1210974. doi: 10.3389/fneur.2023.1210974

Aranda, M. P., Kremer, I. N., Hinton, L., Zissimopoulos, J., Whitmer, R. A., Hummel, C. H., et al. (2021). Impact of dementia: health disparities, population trends, care interventions, and economic costs. *J. Am. Geriatr. Soc.* 69, 1774–1783. doi: 10.1111/jgs.17345

Atkins, A. S., Khan, A., Ulshen, D., Vaughan, A., Balentin, D., Dickerson, H., et al. (2018). Assessment of instrumental activities of daily living in older adults with subjective cognitive decline using the virtual reality functional capacity assessment tool (VRFCAT). *J. Prev Alzheimers Dis.* 5, 216–234. doi: 10.14283/jpad.2018.28

Atkins, A. S., Stroescu, I., Spagnola, N. B., Davis, V. G., Patterson, T. D., Narasimhan, M., et al. (2015). Assessment of age-related differences in functional capacity using the virtual reality functional capacity assessment tool (VRFCAT). *J. Prev Alzheimers Dis.* 2, 121–127. doi: 10.14283/jpad.2015.61

Ault, L., Goubran, R., Wallace, B., Lowden, H., and Knoefel, F. (2020). Smart home technology solution for night-time wandering in persons with dementia. *J. Rehabil. Assist. Technol. Eng.* 7:7. doi: 10.1177/2055668320938591

Bangen, K. J., Jak, A. J., Schiehser, D. M., Delano-Wood, L., Tuminello, E., Han, S. D., et al. (2010). Complex activities of daily living vary by mild cognitive impairment subtype. *J. Int. Neuropsychol. Soc.* 16, 630–639. doi: 10.1017/S1355617710000330

Bernstein, J. P. K., Dorociak, K., Mattek, N., Leese, M., Trapp, C., Beattie, Z., et al. (2022). Unobtrusive, in-home assessment of older adults' everyday activities and health events: associations with cognitive performance over a brief observation period. *Aging Neuropsychol. Cognit.* 29, 781–798. doi: 10.1080/13825585.2021.1917503

Bouchabou, D., Nguyen, S. M., Lohr, C., LeDuc, B., and Kanellos, I. (2021). A survey of human activity recognition in Smart homes based on IoT sensors algorithms: taxonomies, challenges, and opportunities with deep learning. *Sensors* 21:6037. doi: 10.3390/s21186037

Brooke, J. (1986). "SUS: a "quick and dirty" usability scale" in *Usability Evaluation in Industry*. eds. P. W. Jordan, B. Thomas, B. A. Weerdmeester and A. L. McClelland (London: Taylor and Francis)

Coughlan, G., Laczó, J., Hort, J., Minihane, A. M., and Hornberger, M. (2018). Spatial navigation deficits — overlooked cognitive marker for preclinical Alzheimer disease? *Nat. Rev. Neurol.* 14, 496–506. doi: 10.1038/s41582-018-0031-x

Crane, R., Brewer, J., Feldman, C., Kabat-Zinn, J., Santorelli, S., Williams, J., et al. (2017). What defines mindfulness-based programs? The warp and the weft. *Psychol. Med.* 47, 990–999. doi: 10.1017/S0033291716003317

Creswell, J. D., Lindsay, E. K., Villalba, D. K., and Chin, B. (2019). Mindfulness training and physical health: mechanisms and outcomes. *Psychosom. Med.* 81, 224–232. doi: 10.1097/PSY.0000000000000675

Cummings, J., Zhou, Y., Lee, G., Zhong, K., Fonseca, J., and Cheng, F. (2023). Alzheimer's disease drug development pipeline: 2023. *Alzheimers Dement.* 9:e12385. doi: 10.1002/trc2.12385

Cuzzocreo, J. L., Yassa, M. A., Verduzco, G., Honeycutt, N. A., Scott, D. J., and Bassett, S. S. (2009). Effect of handedness on fMRI activation in the medial temporal lobe during an auditory verbal memory task. *Hum. Brain Mapp.* 30, 1271–1278. doi: 10.1002/hbm.20596

Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, J. L., Blennow, K., et al. (2014). Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet Neurol.* 13, 614–629. doi: 10.1016/S1474-4422(14)70090-0

Dubois, B., Hampel, H., Feldman, H. H., Scheltens, P., Aisen, P., Andrieu, S., et al. (2016). Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimers Dement.* 12, 292–323. doi: 10.1016/j.jalz.2016.02.002

European Medicines Agency. Applications for New Human Medicines under Evaluation by the CHMP; (2023). Available at: https://www.ema.europa.eu/en/medicines/medicines-human-use-under-evaluation (Accessed December 12, 2023)

European Medicines Agency (EMA) (2016). Guideline Good Clinical Practice E6(R2). Available at: https://www.ema.europa.eu/en/ich-e6-r2-good-clinical-practice-scientific-guideline#current-version--section (Accessed December 12, 2023)

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Garcia-Constantino, M., Orr, C., Synnott, J., Shewell, C., Ennis, A., Cleland, I., et al. (2021). Design and implementation of a Smart home in a box to monitor the wellbeing of residents with dementia in care homes. *Front. Digit. Health.* 3:798889. doi: 10.3389/fdgth.2021.798889

Gold, M., Amatniek, J., Carrillo, M. C., Cedarbaum, J. M., Hendrix, J. A., Miller, B. B., et al. (2018). Digital technologies as biomarkers, clinical outcomes assessment, and recruitment tools in Alzheimer's disease clinical trials. *Alzheimers Dement (N Y).* 4, 234–242. doi: 10.1016/j.trci.2018.04.003

Han, S. S., White, K., and Cisek, E. (2022). A feasibility study of individuals living at home with Alzheimer's disease and related dementias: utilization of visual mapping assistive technology to enhance quality of life and reduce caregiver burden. *Clin. Interv. Aging* 17, 1885–1892. doi: 10.2147/CIA.S387255

Hayes, T. L., Abendroth, F., Adami, A., Pavel, M., Zitzelberger, T. A., and Kaye, J. A. (2008). Unobtrusive assessment of activity patterns associated with mild cognitive impairment. *Alzheimers Dement.* 4, 395–405. doi: 10.1016/j.jalz.2008.07.004

Hosmer, DW, and Lemeshow, S, *Applied Logistic Regression*, *2nd*. John Wiley and Sons, New York, NY (2000), pp. 160–164

Hutton, J. T., Dippel, R. L., and Loewenson, R. B. (1998). "Functional rating scale for the symptoms of dementia" in *Handbook of Geriatric Assessment*. eds. J. J. Gallo, W. Reichel and L. Andersen (Rockville, MD: Aspen Publishers), 77–80.

Ioulietta, L., Oikonomou Vangelis, P., Lampros, M., Margarita, G., Vasilis, A., Stavropoulos Thanos, G., et al. (2023). RADAR-AD consortium, eliciting brain waves of people with cognitive impairment during meditation exercises using portable electroencephalography in a smart-home environment: a pilot study. *Front. Aging Neurosci.* 15:1167410. doi: 10.3389/fnagi.2023.1167410

Jekel, K., Damian, M., Storf, H., Hausner, L., and Frölich, L. (2016). Development of a proxy-free objective assessment tool of instrumental activities of daily living in mild cognitive impairment using Smart home technologies. *J. Alzheimers Dis.* 52, 509–517. doi: 10.3233/JAD-151054

Jekel, K., Damian, M., Wattmo, C., Hausner, L., Bullock, R., Connelly, P. J., et al. (2015). Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review. *Alzheimers Res. Ther.* 7:17. doi: 10.1186/s13195-015-0099-0

Karakostas, A., König, A., Crispim-Junior, C. F., Bremond, F., Derreumaux, A., Lazarou, I., et al. (2020). A French-Greek cross-site comparison study of the use of automatic video analyses for the assessment of autonomy in dementia patients. *Biosensors (Basel).* 10:103. doi: 10.3390/bios10090103

Kaur, N., Belchior, P., Gelinas, I., and Bier, N. (2016). Critical appraisal of questionnaires to assess functional impairment in individuals with mild cognitive impairment. *Int. Psychogeriatr.* 28, 1425–1439. doi: 10.1017/S104161021600017X

Khalsa, D. S. (2015). Kirtan Kriya Meditation Research Study. Available at: http://www.alzheimersprevention.org/downloadables/White_Paper.pdf (Accessed March 14, 2024)

Kosmidis, M. H., Vlahou, C. H., Panagiotaki, P., and Kiosseoglou, G. (2004). The verbal fluency task in the Greek population: normative data, and clustering and switching strategies. *J. Int. Neuropsychol. Soc.* 10, 164–172. doi: 10.1017/S1355617704102014

Kounti, F., Tsolaki, M., and Kiosseoglou, G. (2006). Functional cognitive assessment scale (FUCAS): a new scale to assess executive cognitive function in daily life activities in patients with dementia and mild cognitive impairment. *Hum. Psychopharmacol.* 21, 305–311. doi: 10.1002/hup.772

Lawson, L., Mc Ardle, R., Wilson, S., Beswick, E., Karimi, R., and Slight, S. P. (2023). Digital endpoints for assessing instrumental activities of daily living in mild cognitive impairment: systematic review. *J. Med. Internet Res.* 25:e45658. doi: 10.2196/45658

Lazarou, Ioulietta, Grammatikopoulou, Margarita, Mpaltadoros, Lampros, Alepopoulos, Vasilis, Nikolopoulos, Spiros, Stavropoulos, Thanos G., et al. Investigating the Impact of a Real Domestic Smart Home Environment in People with Cognitive Impairment as Part of the RADAR-AD Tier 3 Study: A Feasibility Study. 32nd Alzheimer Europe Conference, Bucharest, Romania, pp. 17–19. (2022)

Lundström, J., de Morais, W. O., Menezes, M., Gabrielli, C., Bentes, J., Sant'Anna, A., et al. (2016). "Halmstad intelligent home - capabilities and opportunities" in *Internet of Things Technologies for HealthCare. HealthyIoT 2016. Lecture Notes of the Institute for Computer Sciences, social Informatics and Telecommunications Engineering*. eds. M. Ahmed, S. Begum and W. Raad, vol. *187* (Cham: Springer)

Ma, K., Gao, M., Guida, F. E., and He, R. (2023). Understanding the influencing factors and mechanism of social compensation for Chinese older adults using social media in the context of smart home: a qualitative analysis. *Front. Public Health* 11:1174920. doi: 10.3389/fpubh.2023.1174920

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* 5, 1315–1316, ISSN 1556-0864,. doi: 10.1097/JTO.0b013e3181ec173d

Miller, L. M., Solomon, D. N., Whitlatch, C. J., Hiatt, S. O., Wu, C. Y., Reynolds, C., et al. (2022). The remote assessment and dynamic response program: development of an in-home dementia-related care needs assessment to improve well-being. *Innov. Aging* 6:igac006. doi: 10.1093/geroni/igac006

Mishra, P., Pandey, C. M., Singh, U., Keshri, A., and Sabaretnam, M. (2019). Selection of appropriate statistical methods for data analysis. *Ann Card Anaesth* 22, 297–301. doi: 10.4103/aca.ACA_248_18

Molinuevo, J. L., Rabin, L. A., Amariglio, R., Buckley, R., Dubois, B., Ellis, K. A., et al. (2017). Subjective cognitive decline initiative (SCD-I) working group. Implementation of subjective cognitive decline criteria in research studies. *Alzheimers Dement.* 13, 296–311. doi: 10.1016/j.jalz.2016.09.012

Moyle, W., Murfield, J., and Lion, K. (2021). The effectiveness of Smart home technologies to support the health outcomes of community-dwelling older adults living with dementia: a scoping review. *Int. J. Med. Inform.* 153:104513. doi: 10.1016/j.ijmedinf.2021.104513

Mpaltadoros, L., Stavropoulos, T. G., Lazarou, I., Alepopoulos, V., Nikolopoulos, S., Ganosi, V., et al. "Support2LIVE: Wearable Sensors, Analytics and the WeAreFamily App for Assisted Living with Dementia", 9th Panhellenic Conference on Biomedical Technologies, Thessaloniki, Greece, pp. 9–11; (2021)

Muurling, M., de Boer, C., Kozak, R., Religa, D., Koychev, I., Verheij, H., et al. (2021). Remote monitoring technologies in Alzheimer's disease: design of the RADAR-AD study. *Alzheimers Res. Ther.* 13:89. doi: 10.1186/s13195-021-00825-4

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Ngandu, T., Lehtisalo, J., Korkki, S., Solomon, A., Coley, N., Antikainen, R., et al. (2022). The effect of adherence on cognition in a multidomain lifestyle intervention (FINGER). *Alzheimers Dement* 18, 1325–1334. doi: 10.1002/alz.12492

Osterrieth, P. A. (1944). Le test de copie d'une figure complexe; contribution à l'étude de la perception et de la mémoire [Test of copying a complex figure; contribution to the study of perception and memory]. *Arch. Psychol.* 30, 206–356.

Owens, A. P., Hinds, C., Manyakov, N. V., Stavropoulos, T. G., Lavelle, G., Gove, D., et al. (2020). Selecting remote measurement technologies to optimize assessment of function in early Alzheimer's disease: a case study. *Front. Psych.* 11:582207. doi: 10.3389/fpsyt.2020.582207

Patel, S. H., and Azzam, P. N. (2005). Characterization of N200 and P300: selected studies of the event-related potential. *Int. J. Med. Sci.* 2, 147–154. doi: 10.7150/ijms.2.147

Petersen, J., Austin, D., Mattek, N., and Kaye, J. (2015). Time out-of-home and cognitive, physical, and emotional wellbeing of older adults: a longitudinal mixed effects model. *PLoS One* 10:e0139643. doi: 10.1371/journal.pone.0139643

Petersen, R. C., Roberts, R. O., Knopman, D. S., Boeve, B. F., Geda, Y. E., Ivnik, R. J., et al. (2009). Mild cognitive impairment: ten years later. *Arch. Neurol.* 66, 1447–1455. doi: 10.1001/archneurol.2009.266

Piau, A., Wild, K., Mattek, N., and Kaye, J. (2019). Current state of digital biomarker Technologies for Real-Life, home-based monitoring of cognitive function for mild cognitive impairment to mild Alzheimer disease and implications for clinical care: systematic review. *J. Med. Internet Res.* 21:e12785. doi: 10.2196/12785

Rasmussen, J., and Langerman, H. (2019). Alzheimer's disease - why we need early diagnosis. *Degener Neurol Neuromuscul Dis* 9, 123–130. doi: 10.2147/DNND.S228939

Rey, A. (1964). *L 'Examen Clinique en Psychologie [Clinical Tests in Psychology]*. Paris: Presses Universitaires de France

Rosen, W. G., Mohs, R. C., and Davis, K. L. (1984). A new rating scale for Alzheimer's disease. *Am. J. Psychiatry* 141, 1356–1364. doi: 10.1176/ajp.141.11.1356

Seelye, A., Leese, M. I., Dorociak, K., Bouranis, N., Mattek, N., Sharma, N., et al. (2020). Feasibility of in-home sensor monitoring to detect mild cognitive impairment in aging military veterans: prospective observational study. *JMIR Form Res.* 4:e16371. doi: 10.2196/16371

Seligman, S. C., Giovannetti, T., Sestito, J., and Libon, D. J. (2014). A new approach to the characterization of subtle errors in everyday action: implications for mild cognitive impairment. *Clin. Neuropsychol.* 28, 97–115. doi: 10.1080/13854046.2013.852624

Sikkes, S. A., Lange-de Klerk, E. S., Pijnenburg, Y. A., Gillissen, F., Romkes, R., Knol, D. L., et al. (2012). A new informant-based questionnaire for instrumental activities of daily living in dementia. *Alzheimers Dement.* 8, 536–543. doi: 10.1016/j.jalz.2011.08.006

Sikkes, S., Lange-de Klerk, E. S. M., Pijnenburg, Y. A. L., Scheltens, P., and Uitdehaag, B. M. J. (2009). A systematic review of instrumental activities of daily living scales in dementia: room for improvement. *J. Neurol. Neurosurg. Psychiatry* 80, 7–12. doi: 10.1136/jnnp.2008.155838

Stavropoulos, T. G., Ioulietta, L., Ana, D., Dianne, G., Jean, G., Manyakov, N. V., et al. (2021b). Wearable devices for assessing function in Alzheimer's disease: a European public involvement activity about the features and preferences of patients and caregivers. *Front. Aging Neurosci.* 13:643135. doi: 10.3389/fnagi.2021.643135

Stavropoulos, T. G., Lazarou, I., Alepopoulos, V., Mpaltadoros, L., Grammatikopoulou, M., Muurling, M., et al. "Digital Biomarkers for Alzheimer's Disease in a Real-Living Smart Home Environment: The RADAR-AD tier 3 Study", Alzheimer Europe Conference; (2021a)

Stavropoulos, T. G., Meditskos, G., Briassouli, A., and Kompatsiaris, I. (2016). "Multimodal sensing and intelligent fusion for remote dementia care and support" in *Proceedings of the 2016 ACM Workshop on Multimedia for Personal Health and Health Care (MMHealth'16)* (New York, NY, USA: Association for Computing Machinery), 35–39.

Stavropoulos, T. G., Papastergiou, A., Mpaltadoros, L., Nikolopoulos, S., and Kompatsiaris, I. (2020). IoT wearable sensors and devices in elderly care: a literature review. *Sensors* 20:2826. doi: 10.3390/s20102826

Stucki, R. A., Urwyler, P., Rampa, L., Müri, R., Mosimann, U. P., and Nef, T. (2014). A web-based non-intrusive ambient system to measure and classify activities of daily living. *J. Med. Internet Res.* 16:e175. doi: 10.2196/jmir.3465

Thomas, N. W., Beattie, Z., Riley, T., Hofer, S., and Kaye, J. (2021). Home-based assessment of cognition and health measures: the collaborative aging research using technology (CART) initiative and international collaborations. *IEEE Instrum. Meas. Mag.* 24, 68–78. doi: 10.1109/MIM.2021.9513638

Tiersen, F., Batey, P., Harrison, M. J. C., Naar, L., Serban, A. I., Daniels, S. J. C., et al. (2021). Smart home sensing and monitoring in households with dementia: user-centered design approach. *JMIR aging* 4:e27047. doi: 10.2196/27047

Tombaugh, T. (2004). Trail making test a and B: normative data stratified by age and education. *Arch. Clin. Neuropsychol.* 19, 203–214. doi: 10.1016/S0887-6177(03)00039-8

Urwyler, P., Stucki, R., Rampa, L., Müri, R., Mosimann, U. P., and Nef, T. (2017). Cognitive impairment categorized in community-dwelling older adults with and without dementia using in-home sensors that recognise activities of daily living. *Sci. Rep.* 7:42084. doi: 10.1038/srep42084

van Dyck, C. H., Swanson, C. J., Aisen, P., Bateman, R. J., Chen, C., Gee, M., et al. (2023). Lecanemab in early Alzheimer's disease. *N. Engl. J. Med.* 388, 9–21. doi: 10.1056/NEJMoa2212948

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* 54, 1063–1070. doi: 10.1037/0022-3514.54.6.1063

Wei, W., Gong, X., Li, J., Tian, K., and Xing, K. (2023). A study on community older people's willingness to use smart home-an extended technology acceptance model with intergenerational relationships. *Front. Public Health* 11:1139667. doi: 10.3389/fpubh.2023.1139667

Wilson, B., Cockburn, J., Baddeley, A., and Hiorns, R. (1989). The development and validation of a test battery for detecting and monitoring everyday memory problems. *J. Clin. Exp. Neuropsychol.* 11, 855–870. doi: 10.1080/01688638908400940

World Health Organisation Dementia Key Facts. (2022). Available at: https://www.who.int/news-room/fact-sheets/detail/dementia. (Accessed April 28, 2022)

Yu, J., Huang, J., and Yang, Q. (2023). Long-term adoption or abandonment of Smart Technology in the Chinese Elderly Home Care Environment: a qualitative research study. *Healthcare (Basel, Switzerland)* 11:2440. doi: 10.3390/healthcare11172440

frontiers | Frontiers in Digital Health

TYPE Mini Review
PUBLISHED 04 July 2024
DOI 10.3389/fdgth.2024.1423790

OPEN ACCESS

EDITED BY
Graham Jones,
Tufts Medical Center, United States

REVIEWED BY
Brant Chapman,
Bergen Community College, United States

*CORRESPONDENCE
Rotem Z. Bar-Or
✉ rotem@neuralight.ai

RECEIVED 26 April 2024
ACCEPTED 20 June 2024
PUBLISHED 04 July 2024

CITATION
Band TG, Bar-Or RZ and Ben-Ami E (2024) Advancements in eye movement measurement technologies for assessing neurodegenerative diseases.
Front. Digit. Health 6:1423790.
doi: 10.3389/fdgth.2024.1423790

Advancements in eye movement measurement technologies for assessing neurodegenerative diseases

Tali G. Band, Rotem Z. Bar-Or* and Edmund Ben-Ami

Department of Neuroscience, NeuraLight Ltd., Tel Aviv, Israel

Eye movements have long been recognized as a valuable indicator of neurological conditions, given the intricate involvement of multiple neurological pathways in vision-related processes, including motor and cognitive functions, manifesting in rapid response times. Eye movement abnormalities can indicate neurological condition severity and, in some cases, distinguish between disease phenotypes. With recent strides in imaging sensors and computational power, particularly in machine learning and artificial intelligence, there has been a notable surge in the development of technologies facilitating the extraction and analysis of eye movements to assess neurodegenerative diseases. This mini-review provides an overview of these advancements, emphasizing their potential in offering patient-friendly oculometric measures to aid in assessing patient conditions and progress. By summarizing recent technological innovations and their application in assessing neurodegenerative diseases over the past decades, this review also delves into current trends and future directions in this expanding field.

KEYWORDS

eye tracking, eye movement, neurodegeneration, digital biomarkers, computer vision, machine learning, patient experience

1 Introduction

Neurodegenerative Diseases (NDs) represent a diverse spectrum of conditions characterized by progressive neuronal dysfunction within the Central Nervous System (CNS), potentially culminating in neural cell death. Noteworthy among these disorders are Alzheimer's disease (AD), Parkinson's Disease (PD), and Amyotrophic Lateral Sclerosis (ALS), collectively impacting millions globally each year. For instance, estimates suggest that approximately 4.7 million individuals aged 65 years or older were diagnosed with AD in the United States in 2010 (1), with a global prevalence of 6.1 million for PD in 2016 (2). As the prevalence of NDs correlates with aging demographics (3), projections anticipate a substantial increase in their volume in the coming years.

The clinical presentation of NDs encompasses a wide range of symptoms spanning motor, cognitive, and behavioral domains, exhibiting considerable variability not only between different disorders but also among individuals. This clinical heterogeneity, compounded by the gradual accumulation of symptoms preceding a definitive diagnosis, poses significant challenges to accurate diagnosis. Moreover, many NDs feature a pre-symptomatic phase, which may extend over several years before the onset of apparent clinical symptoms. This pre-symptomatic period holds significant implications for potential preventive interventions and disease-modifying therapies. However, current diagnostic modalities often lack the sensitivity required to detect NDs

during this critical phase. Even after diagnosis, uncertainties persist, particularly in the early stages, further complicated by the diverse clinical spectrum encompassed within Parkinson's plus syndromes (4–6). The inadequacy of reliable diagnostic tools, coupled with the inherent subjectivity of clinical evaluation and inter-individual variability, underscores the urgent need for objective biomarkers capable of capturing both motor and cognitive processes (7, 8).

Abnormalities in eye movements are evident in various NDs, including PD (9–12), ALS (13–16), and AD (17–19). Several types of eye movements are affected by neuronal pathology, of which the most prominent are saccades—rapid eye movements that move the line of sight between successive points of fixation (20). Some of these oculomotor abnormalities cannot be observed during a standard clinical examination, and a recording is required to obtain accurate and objective measurements (7, 16, 21). Various oculomotor abnormalities hold promise as potential biomarkers for both diagnosing NDs and monitoring their progression (7, 10, 16, 21). In numerous studies, oculometric measures (OMs) demonstrate temporal reliability and stay consistent over short intervals (22–24).

Some OMs were shown to distinguish between different phenotypes of indications with similar clinical symptoms, providing valuable insights into disease progression and management. For instance, smooth pursuit eye movement features, and specifically gain, were found to be significantly different in the early stages of PD, Progressive Supranuclear Palsy (PSP), Corticobasal syndrome (CBD), and Multiple System Atrophy (MSA), when many clinical symptoms are similar or indistinctive (6, 12, 18).

Although various OMs strongly correlate with several ND acknowledged outcome measures, the optimal OM compositions, and their interpretation remain contingent on the diagnosed disease. For example, in PD patients, correlations have been observed between scores on the MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and anti-saccade features (25), while ALS patients show similar correlations of anti-saccade latency and error rate with their Revised Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS-R) scores—with additional correlation between smooth pursuit intrusive saccade rate and their ALSFRS-R scores (26, 27). Similarly, patients with AD demonstrate robust correlations between their Mini Mental State Examination (MMSE) scores and specific OMs, namely pro-saccadic latency (28) and micro-saccade lateral bias (19).

In this mini-review, we provide an overview of the current landscape of sensors and techniques for assessing OM abnormalities (Section 2.1) and explore various examination environments and setups facilitating objective oculomotor measurements (Section 2.2). Additionally, we discuss the implications, clinical trends, and anticipated advancements in Section 3.

2 Extracting oculometric measures

Accurate gaze estimations in eye-tracking technology are contingent upon several critical parameters to ensure precision while prioritizing patient comfort and ease of use. When considering sample precision, achieving high spatial resolution and addressing system-specific artifacts to ensure measurement accuracy are emphasized. Simplifying setup complexity in terms of device size and complexity, software interfaces, and operator and user interfaces will be considered when accounting for the patient experience. Striving for a non-invasive and user-friendly setup is essential to minimize patient burden, enabling eye movement measurements without necessitating head restraint or other restrictive measures.

2.1 Sensors and technology

Theoretically, all OMs may be accurately extracted given a long enough time series of the subject gaze direction with sufficient temporal and spatial resolution. Therefore, the main challenges in the sensing system used for OM extraction are spatial accuracy and sample rate, limiting the OM types that may be extracted using a specific sensor. Pro saccadic latency (the time interval between the stimuli of a saccadic eye movement and the beginning of the actual eyeball movement) values can be as short as 100 ms, and the visual angle amplitude of microsaccades may be as narrow as 0.1 degrees (29). Therefore, OM extraction abilities depend on the sensors' combination of spatial accuracy and sample rate.

Along with the physical limitations of OM extraction, the advancements in computation capabilities and the reduction in electronic component sizes enabled the development of sensors that are both accurate, affordable, and patient-friendly, led by video imagers accompanied by dedicated computer vision software. Due to these trends of increasing measurement accuracy and smaller equipment size, sensors are now planned to be used more frequently across diverse populations. Therefore, they are required to be as comfortable as possible, with minimal burden on the subjects being examined (30, 31).

2.1.1 Electrooculography

Electrooculography (EOG) entails the strategic placement of electrodes on the periorbital skin to monitor voltage fluctuations corresponding to ocular movements. This technique facilitates the distinct capture of both horizontal and vertical eye movements, even in the absence of any ocular stimulation and when the subjects' eye may be closed (30, 32). While EOG presents distinct advantages in terms of cost-effectiveness and minimal energy consumption compared to alternative eye-tracking modalities, its resolution remains constrained by susceptibility to artifacts, mainly from external sources such as ambient electrical field perturbations. Moreover, physiological artifacts originating from muscular activity, particularly during bodily motions or facial muscle engagement by participants, contribute further to signal distortion (32–34). Nevertheless, with continued research endeavors, substantial potential exists to augment precision and advance the sophistication of EOG technologies.

2.1.2 Scleral coil system

The scleral coil system involves affixing one or two coils onto the ocular globe and quantifying induced voltages resulting from alterations in magnetic fields precipitated by ocular movements. Most scleral coil sensors are shaped as a ring and placed on the sclera surrounding the iris, similar to eye contacts, but often with an additional wire that connects devices external to the eye. Despite the advantageous feature of minimal noise interference, its invasive nature warrants consideration, as its utilization is constrained to a recommended duration of 30 min or less, primarily due to predictable discomforts, including ocular dryness and transient corneal deformities. It is imperative to acknowledge that the presence of the search coil significantly affects select oculomotor parameters, notably saccadic latency and peak velocity (35, 36).

2.1.3 Video oculography

Video Oculography (VOG), a technology grounded in non-invasive video graphics, has garnered increasing attention over the past two decades (37–39). This innovative approach employs one or multiple cameras (monochromatic or multispectral) to carefully examine the gathered data from captured images, seamlessly adaptable through either a head-mounted apparatus or the integration of head-free webcams within computing systems or handheld devices (30).

The fusion of video-based tracking alongside advanced software platforms for the extraction of OMs has recently demonstrated remarkable precision (16, 31, 40, 41). Vigorous endeavors are underway to develop models to augment accuracy and achieve precision levels commensurate with contemporary IR eye trackers, obviating the necessity for supplementary configurations or costly apparatus. Extensive research continues exploring the efficacy of machine learning and neural network architectures in tracking ocular movements (42). Noteworthy among these advancements is the NeuraLight platform, leveraging the video-based NeuraLight Gaze Estimation Model (NLGEM) and the Calibrated Gaze Model (CGM), showcasing equivalence to established references such as the Tobii eye tracker (43, 44). The CGM model capitalizes on visual stimuli for measurements, facilitating ongoing real-time calibration during testing sessions and obviating the requirement for discrete calibration procedures (45).

2.1.4 Infrared eye tracker

A notable advancement in the domain of eye tracking materialized with the advent of infrared (IR) systems. These setups emit infrared light toward the subject's eyes, measuring the reflected light to ascertain the precise location of the pupil's center and to approximate the Point-of-Regard (PoR, the location the subject is looking at) (46, 47). Such systems can be mounted on desktop monitors and laptops, and even integrated into wearable head devices (44, 48). Initially, achieving a stable head orientation was imperative for ensuring measurement accuracy; however, contemporary iterations of eye trackers demonstrate commendable resilience, swiftly recuperating from head movements. Calibration remains a prerequisite for accurate PoR detection within this framework. Despite exhibiting relatively minimal noise levels in comparison to EOG, this methodology still has its susceptibilities, with various artifacts potentially arising from idiosyncratic patient factors, encompassing eyelid morphology, eyelash length, and the utilization of corrective lenses or spectacles (12, 47, 49). Noteworthy is the treatment of IR technology as a distinct entity within the broader purview of VOG video-based gaze systems (31). Within the confines of this mini-review, we opt to delineate IR technology as a discrete entity, underscored by its distinctive attributes. Unlike the conventional camera-based VOG systems, which passively receive data in the visual light spectrum, relying on external light sources, IR technology proffers an active component, thereby imbuing it with an additional layer of functionality. However, similar to VOG methodologies, inherent challenges may emerge when applying IR tracking technique from diverse patient characteristics.

2.2 Examination setup

In the past, eye-tracking systems necessitated intrusive measures, demanding physical constraints on the subject's head. Typically, studies relied on visual stimuli presented of a singular monitor, screen, or light array, with participants seated, and their heads immobilized using various means such as chin rests. A visual stimulus would manifest on a screen positioned before them, set at a specific visual height and a viewing distance typically ranging between 60 cm and 70 cm (27, 50). However, strides within the field have introduced non-intrusive configurations, allowing participants unrestricted head movements while upholding precision (51). Ambient factors, including light sources, exert notable influence on the accuracy of gaze detection. Optimal laboratory conditions dictate sound and light isolated rooms to mitigate distractions and keep participants focused on the assigned tasks. Hence, minimizing direct and ambient sunlight, oftentimes achieved by dimming or even extinguishing non-essential lighting sources, proves indispensable. A well-lit environment ensures pupil constriction, consequently enhancing data quality, particularly in scenarios involving variable luminance stimuli (52).

The evolution of physical eye-tracking setups has been remarkable, transitioning from conventional head-mounted apparatuses (48) to the integration of webcams within computing devices or handheld gadgets such as smartphones (53) or tablets (54). This progression facilitates precise measurements without the necessity of head restraining. For instance, in tablet setups, the device screen is oriented vertically, with the camera side facing upward, securely affixed at eye level through the employment of a tablet pole mount. The subject face is positioned at an approximate distance of 45 cm from the screen (55). Ongoing endeavors are directed toward refining head-free tracking capabilities accommodating variations in head positioning, distances, and illumination conditions (56). Challenges specific to this domain, such as "head-gaze correlation overfitting" and "head pose ambiguity," are diligently addressed in pursuit of better accuracy.

Empirical evidence suggests that the accuracy of measurements obtained solely through smartphone utilization rivals that of dedicated eye-tracking systems. Notably, one study reported a minimal error margin of 0.46 cm on the smartphone screen (equivalent to 0.6–1 degrees of viewing angle), requiring less than 30 s of calibration data per user (42). The potential for enhancing smartphone-based eye-tracking systems through refined calibration methodologies remains palpable (53). Embracing smartphone technology presents a cost-effective alternative to conventional eye-tracking devices, fostering scalability and enabling broader sample sizes in clinical research endeavors (42).

Looking ahead, advancements in virtual reality (VR) technology promise to broaden the scope of eye-tracking applications. Although VR users, particularly those with neurodegenerative diseases, may experience cybersickness (57, 58) or even oculomotor function changes (59), integrating eye-tracking functionalities into VR headsets adds a spatial dimension to stimuli, enriching the interactive experience (5). These technological strides pave the way for more cost-effective and portable equipment, thereby extending the reach of eye tracking beyond traditional laboratory or clinical settings to telemedical homes and external environments. The expanded accessibility enhances the prospect of engaging a more diverse pool of patients and control groups for comparative analyses (31, 49), thereby fostering the accumulation of richer datasets. With increased data availability, the potential for accuracy enhancement and further technological refinement is substantial.

3 Discussion

The integration of high-precision gaze-detection systems with accessible setups holds promise for significant scientific advancements in analyzing eye movement and oculometric measurements among patients with NDs (45). These advancements facilitate comfortable examination procedures in clinics as well as remote measurements in patients' homes, assuming access to a computing device equipped with a webcam (e.g., tablet, smartphone, laptop, or desktop computer). Such remote monitoring enables the sampling of a vastly larger number of patients, expanding the training sets of various models and thereby enhancing their accuracy. Higher accuracy and reliability are suggested to expand usage and increase the volumes of the training data, closing a positive feedback loop. Moreover, validated video-based remote OM assessment platforms are expected to reduce costs in pharmaceutical clinical trials and accelerate the usage volume growth, adding diverse demographics and ethnicities and providing positive reinforcement for the projected adaptation of these platforms. Looking ahead, self-operated home-based ND monitoring is a feasible emerging stage in the future ND assessment protocol.

As a field adjusting to recently developed technology, further research is required to study the correlations between ND outcome measures and both traditional and newly developed OMs. However, we surf the front wave of the relevant technologies and develop the ND assessment platforms accordingly. Rapid improvements in computational abilities include complex machine learning and artificial intelligence models, as well as the reduction and acceleration of processing units and data storage devices. In addition, hardware recent developments of imaging sensors in various wavelengths and sensitivities, both eye tracking technologies offer an opportunity to integrate their input with the booming computational power and soon finalize a first validated platform for ND assessment based on eye movements.

Although minimal and preferably seamless, advanced examination setups for eye-movement abnormality measurement present certain challenges. Head-mounted devices are limited in sampling rate, particularly affecting saccade analysis; handheld devices encounter numerous issues, including spatial resolution discrepancies among different cameras, introducing variability in head positions and angles, and variations in distance from the camera (49). These challenges must be addressed to enhance the efficacy of eye-tracking methodologies in ND research.

Future trends in eye-movement assessment for NDs may introduce platforms that continuously measure and extract OMs without pre-defined visual stimuli. Such platforms may be integrated with daily used displays like smartphones or desktop computers or on any future computing device that enables eye tracking, including Virtual Reality (VR) or Augmented Reality (AR) devices. VR/AR devices introduce additional dimension to the apparent visual field (depth), enabling the extraction of unique OMs that are influenced by the depth coordinate of the PoR.

Assessing ND severity and progression using eye-movement abnormality measurements and the development and definition of the OMs that will found these measurements are emerging applications in their booming stage. While initial promising results have already been shown in recent software and hardware studies, recent advancements have introduced a feasible potential for a more affordable and patient-friendly platform for assessing ND condition and progression.

Author contributions

TB: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. RB-O: Conceptualization, Project administration, Writing – review & editing. EB-A: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article.

The research was fully funded by NeuraLight Ltd.

Conflict of interest

All authors are employees of NeuraLight Ltd. The manuscript preparation and submission were fully funded by NeuraLight Ltd.

Publisher's note

References

1. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer Disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*. (2013) 80 (19):1778–83. doi: 10.1212/WNL.0b013e31828726f5

2. Dorsey ER, Elbaz A, Nichols E, Abbasi N, Abd-Allah F, Abdelalim A, et al. Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol*. (2018) 17(11):939–53. doi: 10.1016/S1474-4422(18)30295-3

3. Gooch CL, Pracht E, Borenstein AR. The burden of neurological disease in the United States: a summary report and call to action: burden of neurological disease. *Ann Neurol*. (2017) 81(4):479–84. doi: 10.1002/ana.24897

4. Antoniades CA, Kennard C. Ocular motor abnormalities in neurodegenerative disorders. *Eye*. (2015) 29(2):200–7. doi: 10.1038/eye.2014.276

5. Przybyszewski AW, Śledzianowski A, Chudzik A, Szlufik S, Koziorowski D. Machine learning and eye movements give insights into neurodegenerative disease mechanisms. *Sensors*. (2023) 23(4):2145. doi: 10.3390/s23042145

6. Rivaud-Pechoux S, Vidailhet M, Brandel JP, Gaymard B. Mixing pro- and antisaccades in patients with parkinsonian syndromes. *Brain*. (2006) 130(1):256–64. doi: 10.1093/brain/awl315

7. Anderson TJ, MacAskill MR. Eye movements in patients with neurodegenerative disorders. *Nat Rev Neurol*. (2013) 9(2):74–85. doi: 10.1038/nrneurol.2012.273

8. Noseworthy JH, Vandervoort MK, Wong CJ, Ebers GC. Interrater variability with the expanded disability Status scale (EDSS) and functional systems (FS) in a multiple sclerosis clinical trial. *Neurology*. (1990) 40(6):971–971. doi: 10.1212/WNL.40.6.971

9. Li D, Butala AA, Moro-Velazquez L, Meyer T, Oh ES, Motley C, et al. Automating the analysis of eye movement for different neurodegenerative disorders. *Comput Biol Med*. (2024) 170:107951. doi: 10.1016/j.compbiomed.2024.107951

10. Reiner J, Franken L, Raveh E, Rosset I, Kreitman R, Ben-Ami E, et al. Oculometric measures as a tool for assessment of clinical symptoms and severity of Parkinson's disease. *J Neural Transm*. (2023) 130:1241–8. doi: 10.1007/s00702-023-02681-y

11. Jung I, Kim JS. Abnormal eye movements in parkinsonism and movement disorders. *J Mov Disord*. (2019) 12(1):1–13. doi: 10.14802/jmd.18034

12. Pinkhardt EH, Kassubek J. Ocular motor abnormalities in parkinsonian syndromes. *Parkinsonism Relat Disord*. (2011) 17(4):223–30. doi: 10.1016/j.parkreldis.2010.08.004

13. Rojas P, Ramírez AI, Fernández-Albarral JA, López-Cuenca I, Salobrar-García E, Cadena M, et al. Amyotrophic lateral sclerosis: a neurodegenerative motor neuron disease with ocular involvement. *Front Neurosci*. (2020) 14:566858. doi: 10.3389/fnins.2020.566858

14. Donaghy C, Thurtell MJ, Pioro EP, Gibson JM, Leigh RJ. Eye movements in amyotrophic lateral sclerosis and its mimics: a review with illustrative cases. *J Neurol Neurosurg Psychiatry*. (2011) 82(1):110–6. doi: 10.1136/jnnp.2010.212407

15. Sharma R, Hicks S, Berna CM, Kennard C, Talbot K, Turner MR. Oculomotor dysfunction in amyotrophic lateral sclerosis: a comprehensive review. *Arch Neurol*. (2011) 68(7):857. doi: 10.1001/archneurol.2011.130

16. Raveh E, Ben-Shimon A, Anisimov V, Kreitman R, Ben-Ami E, Nechushtan E, et al. Correlation between oculometric measures and clinical assessment in ALS patients participating in a phase IIb clinical drug trial. *Amyotroph Lateral Scler Front Degener*. (2023) 24(5–6):495–501. doi: 10.1080/21678421.2023.2196315

17. Opwonya J, Doan DNT, Kim SG, Kim JI, Ku B, Kim S, et al. Saccadic eye movement in mild cognitive impairment and Alzheimer's disease: a systematic review and meta-analysis. *Neuropsychol Rev*. (2022) 32:193–227. doi: 10.1007/s11065-021-09495-3

18. Garbutt S, Matlin A, Hellmuth J, Schenk AK, Johnson JK, Rosen H, et al. Oculomotor function in frontotemporal lobar degeneration, related disorders and Alzheimer's disease. *Brain*. (2008) 131(5):1268–81. doi: 10.1093/brain/awn047

19. Kapoula Z, Yang Q, Otero-Millan J, Xiao S, Macknik SL, Lang A, et al. Distinctive features of microsaccades in Alzheimer's disease and in mild cognitive impairment. *Age (Omaha)*. (2014) 36(2):535–43. doi: 10.1007/s11357-013-9582-3

20. Leigh RJ, Kennard C. Using saccades as a research tool in the clinical neurosciences. *Brain*. (2004) 127(3):460–77. doi: 10.1093/brain/awh035

21. MacAskill MR, Anderson TJ. Eye movements in neurodegenerative diseases. *Curr Opin Neurol*. (2016) 29(1):61–8. doi: 10.1097/WCO.0000000000000274

22. Iacono WG, Lykken DT. Two-year retest stability of eye tracking performance and a comparison of electro-oculographic and infrared recording techniques: evidence of EEG in the electro-oculogram. *Psychophysiology*. (1981) 18(1):49–55. doi: 10.1111/j.1469-8986.1981.tb01542.x

23. Roy-Byrne P, Radant A, Wingerson D, Cowley DS. Human oculomotor function: reliability and diurnal variation. *Biol Psychiatry*. (1995) 38(2):92–7. doi: 10.1016/0006-3223(94)00225-R

24. Blekher T, Weaver MR, Cai X, Hui S, Marshall J, Jackson JG, et al. Test–retest reliability of saccadic measures in subjects at risk for huntington disease. *Investig Opthalmology Vis Sci*. (2009) 50(12):5707. doi: 10.1167/iovs.09-3538

25. Antoniades CA, Demeyere N, Kennard C, Humphreys GW, Hu MT. Antisaccades and executive dysfunction in early drug-naive Parkinson's disease: the discovery study: ANTISACCADES IN EARLY PD. *Mov Disord*. (2015) 30(6):843–7. doi: 10.1002/mds.26134

26. Donaghy C, Pinnock R, Abrahams S, Cardwell C, Hardiman O, Patterson V, et al. Slow saccades in bulbar-onset motor neurone disease. *J Neurol*. (2010) 257 (7):1134–40. doi: 10.1007/s00415-010-5478-7

27. Marti-Fàbregas J, Roig C. Oculomotor abnormalities in motor neuron disease. *J Neurol*. (1993) 240(8):475–8. doi: 10.1007/BF00874116

28. Yang Q, Wang T, Su N, Xiao S, Kapoula Z. Specific saccade deficits in patients with Alzheimer's disease at mild to moderate stage and in patients with amnestic mild cognitive impairment. *Age (Omaha)*. (2013) 35(4):1287–98. doi: 10.1007/s11357-012-9420-z

29. Krauzlis RJ. Eye movements. In: Squire L, Berg D, Bloom FE, du Lac S, Ghosh A, Spitzer NC, editors. *Fundamental Neuroscience*. Waltham, MA: Elsevier (2013). p. 697–714. Available online at: https://linkinghub.elsevier.com/retrieve/pii/B9780123858702000329 (cited April 24, 2024).

30. Chennamma HR, Yuan X. A Survey on Eye-Gaze Tracking Techniques. arXiv. (2013). Available online at: http://arxiv.org/abs/1312.6410 (cited December 10, 2023)m

31. Kullmann A, Ashmore RC, Braverman A, Mazur C, Snapp H, Williams E, et al. Portable eye-tracking as a reliable assessment of oculomotor, cognitive and reaction time function: normative data for 18–45 year old. *PLoS One*. (2021) 16(11): e0260351. doi: 10.1371/journal.pone.0260351

32. Sun YR, Beylergil SB, Gupta P, Ghasia FF, Shaikh AG. Monitoring eye movement in patients with Parkinson's disease: what can it tell US? *Eye Brain*. (2023) 15:101–12. doi: 10.2147/EB.S384763

33. Corin MS, Elizan TS, Bender MB. Oculomotor function in patients with Parkinson's disease. *J Neurol Sci*. (1972) 15(3):251–65. doi: 10.1016/0022-510X(72)90068-8

34. Belkhiria C, Peysakhovich V. Electro-encephalography and electro-oculography in aeronautics: a review over the last decade (2010–2020). *Front Neuroergonomics*. (2020) 1:606719. doi: 10.3389/fnrgo.2020.606719

35. Geest JN VD, Frens MA. Recording eye movements with video-oculography and scleral search coils: a direct comparison of two methods. *J Neurosci Methods*. (2002) 114(2):185–95. doi: 10.1016/S0165-0270(01)00527-1

36. Sprenger A, Neppert B, Köster S, Gais S, Kömpf D, Helmchen C, et al. Long-term eye movement recordings with a scleral search coil-eyelid protection device allows new applications. *J Neurosci Methods*. (2008) 170(2):305–9. doi: 10.1016/j.jneumeth.2008.01.021

37. Kim SC, Nam KC, Lee WS, Kim DW. A new method for accurate and fast measurement of 3D eye movements. *Med Eng Phys*. (2006) 28(1):82–9. doi: 10.1016/j.medengphy.2005.04.002

38. Schreiber K, Haslwanter T. Improving calibration of 3-D video oculography systems. *IEEE Trans Biomed Eng*. (2004) 51(4):676–9. doi: 10.1109/TBME.2003.821025

39. Sakatani T, Isa T. PC-based high-speed video-oculography for measuring rapid eye movements in mice. *Neurosci Res*. (2004) 49(1):123–31. doi: 10.1016/j.neures.2004.02.002

40. Imai T, Sekine K, Hattori K, Takeda N, Koizuka I, Nakamae K, et al. Comparing the accuracy of video-oculography and the scleral search coil system in human eye movement analysis. *Auris Nasus Larynx*. (2005) 32(1):3–9. doi: 10.1016/j.anl.2004.11.009

41. Rosset I, Raveh E, Shimon AB, Anisimov V, Ben-Ami E, Kreitman R, et al. Validation of a novel software-based platform to extract oculometric

measures. *Acta Ophthalmol (Copenh)*. (2022) 100(S275):0359. doi: 10.1111/j.1755-3768.2022.0359

42. Valliappan N, Dai N, Steinberg E, He J, Rogers K, Ramachandran V, et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nat Commun*. (2020) 11(1):4553. doi: 10.1038/s41467-020-18360-5

43. Housholder A, Reaban J, Peregrino A, Votta G, Mohd TK. Evaluating accuracy of the tobii eye tracker 5. In: Kim JH, Singh M, Khan J, Tiwary US, Sur M, Singh D, editors. *Intelligent Human Computer Interaction*. Cham: Springer International Publishing (2022). p. 379–90. (Lecture Notes in Computer Science; vol. 13184). Available online at: https://link.springer.com/10.1007/978-3-030-98404-5_36 (cited December 24, 2023).

44. Onkhar V, Dodou D, De Winter JCF. Evaluating the tobii pro glasses 2 and 3 in static and dynamic conditions. *Behav Res Methods*. (2023). doi: 10.3758/s13428-023-02173-7. [Epub ahead of print]

45. Harpaz E, Bar-Or RZ, Rosset I, Ben-Ami E. Video-based gaze detection for oculomotor abnormality measurements. *Appl Sci*. (2024) 14(4):1519. doi: 10.3390/app14041519

46. Guestrin ED, Eizenman M. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans Biomed Eng*. (2006) 53(6):1124–33. doi: 10.1109/TBME.2005.863952

47. Bueno APA, Sato JR, Hornberger M. Eye tracking—the overlooked method to measure cognition in neurodegeneration? *Neuropsychologia*. (2019) 133:107191. doi: 10.1016/j.neuropsychologia.2019.107191

48. Stein N, Niehorster DC, Watson T, Steinicke F, Rifai K, Wahl S, et al. A comparison of eye tracking latencies among several commercial head-mounted displays. *Percept*. (2021) 12(1):204166952098333. doi: 10.1177/2041669520983338

49. Blignaut P, Wium D. Eye-tracking data quality as affected by ethnicity and experimental design. *Behav Res Methods*. (2014) 46(1):67–80. doi: 10.3758/s13428-013-0343-0

50. Terao Y, Fukuda H, Yugeta A, Hikosaka O, Nomura Y, Segawa M, et al. Initiation and inhibitory control of saccades with the progression of Parkinson's disease—changes in three major drives converging on the superior colliculus. *Neuropsychologia*. (2011) 49(7):1794–806. doi: 10.1016/j.neuropsychologia.2011.03.002

51. Hansen DW, Ji Q. In the eye of the beholder: a survey of models for eyes and gaze. *IEEE Trans Pattern Anal Mach Intell*. (2010) 32(3):478–500. doi: 10.1109/TPAMI.2009.30

52. Holmqvist K. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford; New York: Oxford University Press (2011). p. 537.

53. Parker TM, Badihian S, Hassoon A, Saber Tehrani AS, Farrell N, Newman-Toker DE, et al. Eye and head movement recordings using smartphones for telemedicine applications: measurements of accuracy and precision. *Front Neurol*. (2022) 13:789581. doi: 10.3389/fneur.2022.789581

54. de Villers-Sidani É, Voss P, Guitton D, Cisneros-Franco JM, Ducharme S. A novel tablet-based software for the acquisition and analysis of gaze and eye movement parameters: a preliminary validation study in Parkinson's disease. *Front Neurol*. (2023) 14:1204733. doi: 10.3389/fneur.2023.1204733

55. Koch NA, Voss P, Cisneros-Franco JM, Drouin-Picaro A, Tounkara F, Ducharme S, et al. Eye movement function captured via an electronic tablet informs on cognition and disease severity in Parkinson's disease. *Sci Rep*. (2024) 14(1):9082. doi: 10.1038/s41598-024-59750-9

56. Zhu W, Deng H. *Monocular free-head 3D gaze tracking with deep learning and geometry constraints*. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017).

57. Saredakis D, Szpak A, Birckhead B, Keage HAD, Rizzo A, Loetscher T. Factors associated with virtual reality sickness in head-mounted displays: a systematic review and meta-analysis. *Front Hum Neurosci*. (2020) 14:96. doi: 10.3389/fnhum.2020.00096

58. Kourtesis P, Collina S, Doumas LAA, MacPherson SE. Technological competence is a Pre-condition for effective implementation of virtual reality head mounted displays in human neuroscience: a technological review and meta-analysis. *Front Hum Neurosci*. (2019) 13:342. doi: 10.3389/fnhum.2019.00342

59. Alhassan M, Alhamad F, Bokhary K, Almustanyir A. Effects of virtual reality head-mounted displays on oculomotor functions. *Int J Ophthalmol Vis Sci*. (2021) 6(1):10. doi: 10.11648/j.ijovs.20210601.12

frontiers | Frontiers in Digital Health

TYPE Review
PUBLISHED 08 July 2024
DOI 10.3389/fdgth.2024.1335289

OPEN ACCESS

EDITED BY
Graham Jones,
Tufts Medical Center, United States

REVIEWED BY
Meysam Asgari,
Oregon Health and Science University, United States
Brant Chapman,
Bergen Community College, United States

*CORRESPONDENCE
Di Jiang
✉ di.jiang@cnrc-nrc.gc.ca

RECEIVED 08 November 2023
ACCEPTED 20 June 2024
PUBLISHED 08 July 2024

CITATION
Jiang D, Yan L and Mayrand F (2024) Emotion expressions and cognitive impairments in the elderly: review of the contactless detection approach.
Front. Digit. Health 6:1335289.
doi: 10.3389/fdgth.2024.1335289

Emotion expressions and cognitive impairments in the elderly: review of the contactless detection approach

Di Jiang[1]*, Luowei Yan[2] and Florence Mayrand[2]

[1]Medical Devices Research Centre, National Research Council of Canada, Boucherville, QC, Canada, [2]Department of Psychology, McGill University, Montreal, QC, Canada

The aging population in Canada has been increasing continuously throughout the past decades. Amongst this demographic, around 11% suffer from some form of cognitive decline. While diagnosis through traditional means (i.e., Magnetic Resonance Imagings (MRIs), positron emission tomography (PET) scans, cognitive assessments, etc.) has been successful at detecting this decline, there remains unexplored measures of cognitive health that could reduce stress and cost for the elderly population, including approaches for early detection and preventive methods. Such efforts could additionally contribute to reducing the pressure and stress on the Canadian healthcare system, as well as improve the quality of life of the elderly population. Previous evidence has demonstrated emotional facial expressions being altered in individuals with various cognitive conditions such as dementias, mild cognitive impairment, and geriatric depression. This review highlights the commonalities among these cognitive health conditions, and research behind the contactless assessment methods to monitor the health and cognitive well-being of the elderly population through emotion expression. The contactless detection approach covered by this review includes automated facial expression analysis (AFEA), electroencephalogram (EEG) technologies and heart rate variability (HRV). In conclusion, a discussion of the potentials of the existing technologies and future direction of a novel assessment design through fusion of AFEA, EEG and HRV measures to increase detection of cognitive decline in a contactless and remote manner will be presented.

KEYWORDS

cognitive decline, remote health, contactless detection, machine learning, elderly population

1 Introduction

The cognitive health of the elderly population has grown to be a central issue in our society. Statistics estimated that at least 6.5 million of Americans aged 65 years or over are living with Alzheimer's disease (AD) (1). In Canada, 597,300 individuals were living with dementia in 2020, and this number was projected to reach close to a million by 2030 (2). The demand and reliance on valid and precise diagnostic tools have therefore increased exponentially. Historically, traditional tools such as neuropsychological tests and brain imaging techniques have been the state-of-the-art diagnostic methods and assessment of severity. Although accurate, these techniques involve intense patient participation in the case of tests or intrusive manipulations in the case of MRIs and PET scans. During the same period where AD cases increased, the proportion of elders living in collective

dwellings or assisted living facilities, such as a nursing home, a chronic care facility or a residence for seniors, also evolved significantly. The 2011 census indicated that 7.9% of seniors aged 65 or over resided in a collective dwelling, whereas the 2021 census revealed 28% of those aged 80 and above living in such arrangements (3, 4). Specifically with the impact of COVID-19 pandemic, one in every twenty Canadians aged 65 or over were living in these facilities in 2021 (5, 6). Thus, our healthcare systems are facing an unprecedented situation with continuously increasing needs and burdens. The emerging trend of regrouping of patients in the facilities brings on new possibilities regarding the assessments of their disorders and disabilities, such that this environment could serve as both the treatment and the diagnosis method. For instance, remote and contactless tools could easily be integrated into the living installations which AD patients utilize daily. To this end, existing evidence has demonstrated emotional facial expressions being altered in individuals with various cognitive conditions such as dementia, mild cognitive impairment, and geriatric depression. Technologies such as Automated Facial Expressions Analysis (AFEA) and remote photoplethysmography (rPPG) have been shown to provide accurate and reliable measures which can be related to cognitive health and disease progression. Given that these technologies can be added to daily protocols already administered to patients in care facilities via camera recordings, assessment of patients' health could be completely re-invented such that intrusive methods will be on need-basis and less required, and preliminary diagnosis can occur in community. In this paper, we will review the use of these technologies in the context of various cognitive conditions to enhance the accessibility of treatment and progress tracking for the elderly in a remote and contactless manner.

2 Cognitive impairments in the elderly population

2.1 Dementia and Alzheimer's disease

Dementia, the most well-known disorder associated with the elderly, is a general term for several diseases, including AD (7). Over 350 Canadians on average were diagnosed with dementia every day in 2022 (2). All these data demonstrate that dementia, with its prevalence, is a non-negligible condition in the health assessment of elders in long-term care facilities. Dementia is understood to affect memory, increase confusion, apathy/depression, and leads to a loss of ability to complete everyday tasks (8). Typically, dementia is assessed through various cognitive and neuropsychological tests such as the Mini-Mental Status Examination (MMSE). Brain scans, such as MRIs and PET scans, can also be used to detect dementia through changes in the brain structure, but these are associated with a high cost and demand extensive resources. While it remains exploratory whether the expression of emotions differs between older adults with dementia and healthy ones, several studies in this space provided promising results. For example, studies looking into facial expression of pain found that participants with dementia expressed more pain in their faces than participants in the control group (9). A more recent study also found that across their participant pool from AD research centers, it appeared that dementia patients facially expressed fewer positive emotions during emotion-eliciting events and instead used more negative expressions (10). In addition, AD patients demonstrated an overall increase in facial expressiveness (11). Similarly, other studies have found altered zygomatic activity (i.e., muscles that control smiling) in patients with dementia while viewing emotion-eliciting images when compared to healthy elderly counterparts (12). The flexibility in emotion expressions was also found to be reduced for AD patients, such that they struggled to amplify positive emotions facially (13). That being said, while these techniques have been used increasingly in the clinical world, no automated assessment of pain through facial expressions has been tested as a valid tool for detecting dementia (14), and little effort has been put into relating facial expression analysis to other physiological measures of dementia. Hence, the automation of facial expression analysis, paired with other measures, would therefore provide an interesting option to both detect dementia, as well as monitor it once it is diagnosed.

2.2 Mild cognitive impairment

Dementia is often first diagnosed as mild cognitive impairment (MCI), which makes it one of the first observable conditions and symptoms in one's cognitive decline. MCI is characterized by a limbo state between normal aging and dementia (15). For people with MCI, typical symptoms include memory deficits as well as other reduced cognitive functions that do not hinder or only slightly affect one's instrumental functional abilities. The prevalence of MCI increases with age, with 10.88% of community-dwellers aged 50–59 years and 21.27% of those aged 80 years and above, as indicated by a recent worldwide meta-analysis (16). More importantly, up to 30% of adults who develop MCI will go on to be diagnosed with some form of dementia; typically AD for those who experience memory deficits (17). Early detection of MCI is crucial in reducing one's risk of developing dementia. Traditional detection tools, however, are often targeted towards the impairments found in AD and therefore not very accurate at detecting MCI. In fact, the MMSE, which is one of the most commonly used cognitive tests, can only detect around 18% of MCI cases (18), and it does not provide substantial support for the early detection of dementia in MCI patients (19). Therefore, more tools are necessary to better understand MCI and help early treatment of cognitive decline. It is known that people with cognitive impairments express emotions differently through their faces compared to healthy adults of the same age (20, 21), and a recent effort has been made to use non-invasive, readily available technologies to assess MCI. For example, Fei and colleagues (18, 22) proposed computer vision techniques for the detection of cognitive impairment, including MCI, in the elderly by analyzing facial features. As manual coding of these expressions can be tedious, an automated way of facial expression analysis (i.e., AFEA) could

potentially provide an efficient, contactless, and non-intrusive detection tool for MCI and allow for better prevention of dementia.

2.3 Depression

Often overlooked, depression is one of the most common symptoms in dementia and MCI patients. The prevalence of depression among elderly individuals tends to vary across investigations due to different experimental designs (23–25). For instance, estimates suggest that depression could affect up to 5% of the elderly population and close to 44% in elders requiring residential health care (26, 27). Despite its prevalence, late onset depression remains underdiagnosed and characterized as a part of normal aging. However, depression has serious impacts on the elderly's cognitive and physical health. Late onset of depression can lead to serious cognitive deficits, often similar to those seen in MCI (28). Research has shown that MCI patients are more likely to develop depression, with a prevalence rate between 16.9% and 55% (29). In fact, half of those who experience depression after the age of 65 and along with cognitive impairment will go on to develop AD or other types of dementia. The comorbidity of depression in patients with dementia can vary between 9 and 68% (30). Hence, depression is both seen as a risk factor for dementia as well as a symptom of the disorder. The cognitive damage due to depression can however be reversed before one progresses into dementia but is too often ignored or undiagnosed. Typical assessments of depression such as the Geriatric Depression Scale (GDS) can be misleading and often wrongly diagnose cognitively impaired adults as depressive (31). New methods of detection are therefore needed, one of which could be the analysis of facial expressions or muscle activity. Facial expression analysis is a common tool used in adults with depression. For example, it has been shown that depressed individuals have a loss of facial muscle tone around the mouth, but higher tone in the brow area, which can be associated with anxiety and anger. Overall, depressed individuals express fewer smiles than healthy adults (32). Depressed individuals also demonstrate decreased activities in the cheek and brow areas upon viewing happy and sad images, compared to non-depressed individuals (33). Correlations between depressive symptoms and end-lip, mouth width, mid-top lip, eye-opening, and mid-eyebrow measures have been found in some studies (34), as well as facial indicators of excess activity in the grief regions of the face, even during joy-inducing stimuli (35). Facial expression analysis is able to identify all these small facial changes during expression of emotions, but it has never been specifically applied to geriatric or late onset depression. Therefore, it would be informative to explore the application of facial expression analysis in the elderly for the prognosis and detection of depression, which in turn could contribute to the monitoring of dementia symptoms and progress in the same population.

Here, we outlined 3 different cognitive and affective conditions present in the elderly population related to cognitive impairment (see Table 1 for summary of reviewed articles). While established diagnostic methods have been successful in identifying these diseases and disorders, recent breakthroughs in contactless technologies show that facial movements during affective states can be used to monitor cognitive decline and the severity of conditions. The implications of such methods are not negligible; remote and contactless assessment would allow for frequent updates on the cognitive health of elders at risk without invasive procedures and at a relatively low cost. Furthermore, the addition of such technologies can easily be integrated into care facilities, which house hundreds of patients in one place. This facilitates the routine assessment of cognitive decline daily, fostering a proactive approach instead of relying on periodic assessments that could lead to significant deteriorations of disorders.

3 Automated facial expression analysis, EEG and rPPG in emotion recognition

In recent years, contactless detection for facial expression analysis and emotion recognition has become a growing field, with more interest in its applications in the medical health domain. Today, various automated methods for emotion assessment have been developed to increase the accuracy of emotions through different means. Among them, Facial Expression Analysis, EEG, and heart rate monitoring have been emerging as viable ways to understand human emotions. Both Facial Expression Analysis and heart rate monitoring have been made available through contactless and remote means such as AFEAs and rPPG, respectively. In this section, we will present a summary of these technologies together with their respective accuracy and usability, as well as express the need for joint usage of these methods in emotion detection.

3.1 Facial action coding system and automated facial expression analysis (AFEA)

The Facial Action Coding System (FACS) is a taxonomy system to identify and classify facial movements during expressions of emotions (36). FACS has been used by psychologists for decades and has recently been applied in animations (37, 38). To classify certain facial expressions, FACS uses Action Units (AUs) to pair together different movements by facial muscles (39). A total of 46 main action units makes up FACS, through which 7 emotions can be detected: happiness/joy, sadness, surprise, fear, anger, disgust, and contempt. Traditionally, FACS required coding of AUs by human coders. The training required to become a certified FACS coder is lengthy, with over hundreds of hours spent coding (40). In the last decade, amazing efforts have been made to automate FACS coding to speed up the process and alleviate human efforts. Through deep learning networks, algorithms have been able to successfully track facial movements and AUs, and subsequent emotion classifications (40–43). Analyses on the accuracy of these algorithms have varied, with some reaching nearly 90% accuracy while others fail to reach 50% accuracy (44–46). For this reason, there are many different

TABLE 1 Summary of previous research on cognitive decline and depression and facial mobility.

Citation	Article Type	Sample	Measures included	Stimuli	Main findings
Kunz et al. (9)	Empirical	42 demented elders 54 healthy elders	Facial action coding system	Pressure stimulation	• Facial responses were significantly increased in demented patients compared to healthy controls. • Facial responses were closely related to the intensity of stimulation for demented patients.
Jiang et al. (10)	Empirical	258 healthy elders 235 AD elders	Vision-based facial expression recognition	Images of object scenes	• AD participants expressed significantly fewer positive emotions, more negative emotions, and higher facial expressiveness. • Facial emotions expressed during the test allowed effective differentiation of AD from healthy participant.
Seidl et al. (11)	Empirical	47 AD patients	Facial action coding system	International affective picture system	• Cognitive decline was related to increased facial expressiveness. • Apathetic symptoms appear to be specifically associated with facial expression in AD.
Burton & Kaszniak (12)	Empirical	13 elders with AD 21 healthy elders	Corrugator and zygomatic electromyography (EMG)	International affective picture system	• Change in zygomatic activity was significantly different between AD and healthy groups, with AD subjects demonstrating an inverted pattern of activity compared to controls.
Henry et al. (13)	Empirical	20 healthy elders 20 AD elders	Expressive emotion behavior coding system	Neutral and amusing video clips	• AD is associated with subtle changes in emotion-expressive behavior. • AD group displayed significantly lower positive affect compared with the control group.
Chen et al. (20)	Empirical	99 patients with frontotemporal dementia (FTD) 45 AD patients 37 healthy controls	Subjective emotional experience	Film clips	• Patients with AD and FTD tended to experience more "mixed emotions" when watching emotionally arousing film clips. • FTD patients reported more positive and negative non-target emotions, whereas AD patients reported more positive non-target emotions.
Pressman et al. (21)	Empirical	36 healthy adults 89 patients with a neurodegenerative disease	Expressive emotion behavior coding system	Three short films	• Participants with FTD tended to express less emotion on their faces than they did through self-report. • Differences within diagnostic subgroups.
Fei et al. (18)	Review	N/A	Facial features analysis; Facial features classification	N/A	• Automatic facial expression analysis has the potential to be used for cognitive impairment detection in the elderly. • May be better to use a local method of facial components alignment, employ static approaches in facial feature extraction and facial feature classification.
Fei et al. (22)	Empirical	61 healthy and cognitively impaired elders	Deep neural network-based emotion analysis system	KDEF dataset; Chinese adults dataset; Chinese elderly people dataset	• The classifier was able to detect the cognitive impairment based on the emotion data from the testing dataset with a detection accuracy of 73.3%.
Katsikitis and Pilowsky (32)	Empirical	21 Parkinson's disease patients 20 depressed patients 12 healthy adults	Facial Expression Measurement program	12 humorous cartoons	• Depressed patients shown smaller mid-eyebrow measures compared to the control group. • Depressed and parkinsonian group had significantly less smiles.
Gehricke and Shapiro (33)	Empirical	11 depressed patients 11 healthy adults	Facial EMG	Imagery situations	• Facial muscle activity over the brow and cheek region was reduced in depressed compared to healthy patients during happy and sad imagery. • Lack of social context differences in frowning may suggest social disengagement and an inhibition of sad facial expression.
Stolicyn et al. (34)	Empirical	48 depressed participants	Facial action coding system	Delayed match to sample task; Rapid detection task; Affective distractions	• Symptomatic participants were characterised by less intense mouth and eyelid movements. • Classification accuracy using cross-validation (within-study replication) reached 79%.
Greden et al. (35)	Empirical	29 healthy controls 37 depressed adults	Facial EMG	Imagery situations	• Patients with endogenous depression had EMG levels that differentiated them from healthy subjects. • Depressed participants had significantly greater activity in corrugator happy and corrugator sad imagery trials.

algorithms available that use different methods to develop their AFEA using FACS. Certain technology companies have created "ready-to-use" platforms that can serve multiple usage and provide AFEA to a wide range of professionals. Such products, like iMotions's Affectiva and Noldus' FaceReader, allow for AFEA to occur with video recordings and without the input of

human coders. These platforms all operate under the same rules and mostly use similar algorithms to classify emotion expressions (45). When using these algorithms, the choice of camera hardware to record the data is important, as the resolution will factor into the facial feature detection accuracy. Studies show that cameras that have stable framerates, auto-focus, and allow access to aperture, brightness, and white balance settings offer the best results (47).[1] The Microsoft Kinect RGB-D camera was also found to accurately locate facial features with high resolution (48–51). However, detailed specifications on the appropriate hardware requirements have not been well established.

From a software algorithm perspective, most deep learning networks utilize the Viola-Jones algorithm to detect the presence of faces within an image or video. The Viola-Jones algorithm works by first selecting Haar-like features in images (52). It then creates an integral image and goes through a machine learning algorithm that identifies the best features to detect a face by creating classifiers. Based on which classifiers work the best on training datasets with faces, the best performing ones are kept and then used to discard non-faces in images through a cascade of classifiers. In the last stage, an image is finally classified as a human face. Upon successful face identification, platforms like FaceReader make a 3D model of the face using the Active Appearance Method (AAM) (53). The AAM can locate 500 points on the face and also analyze texture. Based on the location of these points, the AAM can classify facial expressions through the training of the algorithm with over 10,000 images of faces. Once an expression is classified, these platforms can assess the valence and arousal of the expression as well as the intensity of all AUs involved during the expression (53).

Such models and platforms, while having clear advantages and benefits of not needing any pre-programming, require commercial licenses that involve regular payments. In addition, studies have demonstrated their limited suitability to applications. Because they are already pre-trained with some generic datasets of face images, some biases were observed in specific populations (54). While somewhat accurate at detecting AUs in the general adult Caucasian population, some research has found that the accuracy of these models drops significantly when applied to other ethnicities and different age groups [(44, 54, 55), but see (56) for new technology addressing AI bias of skin tone]. Therefore, their usage cannot be applied universally.

Nonetheless, the core foundations of these platforms remain unbiased prior to the training of the algorithms. Independent implementation of a similar platform can be done by utilizing open access deep learning networks. Through the training of the network, a platform could hypothetically be applied to any specific group and obtain accurate readings of facial expression. The challenge, however, consistent with those of most AI/ML algorithms, is the need for large-volume, diverse, and well-representing datasets, which are known to be rare and limited (57). To successfully train an AFEA, thousands of images need to be presented in training to develop highly reliable classifiers. Without such training, the algorithm's accuracy will drop significantly, if not be nonexistent. Furthermore, to apply an AFEA to the aging population to detect conditions such as dementia and Parkinson's disease, an extensive collection of images of elderly people's faces would be necessary to train the algorithm. However, because these clinical populations are less prevalent relative to healthy populations, very few datasets are available (58). Among these few are the University of Regina's Pain in Severe Dementia dataset and the UNBC-McMaster Shoulder Pain Expression Archive dataset. Other larger datasets such as the FACES dataset contain a subcategory with older adults but cannot be used on its own (57). Despite the individual limitations, these smaller datasets could potentially be grouped together to train an algorithm to work on the elderly population. Interesting alternatives were explored by researchers responding to the scarcity of available datasets. For example, online videos, such as YouTube videos, with people involved with Parkinson's disease were used to train the AFEA to recognize patterns of the disease without having to develop their own dataset (59). This proved to be a promising training technique, with an accuracy of over 82% for the detection of Parkinson's disease reported. Such a method could be used on all populations that are underrepresented in large datasets (59). Therefore, the biases seen in most algorithms can be minimized through re-training using various databases and available images/videos.

3.2 EEG and emotion recognition

In parallel to the externally observable and accessible factors of the facial mobility approach to cognitive assessment, measurement and understanding of patients' internal brain activity using EEG data has been considered often as a reference information for clinical evaluation. For this purpose, EEG has been extensively studied in different populations exhibiting cognitive decline as well as in demented patients (e.g., AD patients; see Table 2 for summary of reviewed articles). As a result, there has been a growing consensus within the scientific community regarding the overall significance of this approach.

In resting-state EEG recordings, AD and MCI patients showed an increased spectral power and functional connectivity in the theta and delta bands, which are the slower frequencies of the spectrum (60). Interestingly, participants in the control group showed a decreased spectral power in these bands with advancing age, thus indicating an inverse aging pattern in the AD and MCI groups. AD participants also showed a decreased spectral power and functional connectivity in the alpha band normally observed in healthy aging. Meghdadi et al. (60) also reported that a [theta/alpha] ratio was very good at discriminating AD from MCI and controls, as exhibiting higher values was associated with increasing cognitive impairment and disease progression. Similarly, early-onset AD patients exhibited higher spectral power in the lower frequencies as well as lower spectral power in higher frequencies when compared to age-matched healthy individuals (61). High

[1]The Logitech HD Pro webcam C920 seems to obtain the best results amongst webcams (45).

TABLE 2 Summary of previous research on cognitive decline and EEG.

Citation	Article type	Sample	Measures included	Stimuli	Main findings
Meghdadi et al. (60)	Empirical	26 AD patients 53 MCI patients 246 healthy controls	Resting state EEG	N/A	• Increases in both spectral power and coherence at slower frequencies in ADs and MCIs. • Decreases in spectral power at slower frequencies in HCs with advancing age. • Decreases in the spectral power and coherence at Alpha frequency in ADs, but not in MCIs. • Theta-to-alpha ratio demonstrated the largest and most significant differences between ADs and HCs.
Özbek et al. (61)	Empirical	47 early-onset AD patients (EOAD) 51 late-onset AD patients (LOAD) 49 young healthy controls 51 old healthy controls	Resting state EEG	N/A	• Increases in slow frequency bands and decreases in fast frequency bands in EOADs. • Frontal theta-to-alpha ratio best discriminated between EOADs and young HCs. • More widespread and severe electrophysiological abnormalities in EOADs than LOADs and HCs.
Gaubert et al. (62)	Empirical	314 preclinical AD adults	Resting state EEG	N/A	• Increases in high frequency oscillations and decreases in low frequency oscillations in frontocentral regions. • Different EEG patterns modulated by the degree of amyloid burden.
Palmiero et al. (63)	Mini-review	N/A	Various	N/A	• Contradictory and mixed results. • Increases in left prefrontal EEG activity for approach-related and positive emotions. • Increases in right EEG prefrontal activity for withdrawal-related and negative emotions.
Kisley et al. (64)	Empirical	51 healthy adults	Tasked-evoked EEG	Emotion-eliciting images	• Overall larger LPP amplitudes elicited by negative than by positive images. • Linear decline of LPP amplitudes with advancing age towards negative images • Responses towards positive images remained age invariant
Tsolaki et al. (65)	Empirical	11 young adults 11 elderly adults	Tasked-evoked EEG	Photographs of fear and anger facial expressions	• Larger amplitudes of the N170 early component in elderly adults than in young adults. • Less differentiation of N170 topographic maps between the two negative stimuli in elderly than in young adults. • More differentiation of topographic maps between the age groups in 'anger' than in 'fear'.
Güntekin et al. (66)	Empirical	30 healthy controls 30 AD patients	Tasked-evoked EEG	Photographs of angry, happy, and neutral facial expressions	• HCs: increased Theta power towards angry expressions, and increased right hemispheric alpha power. • ADs decreased Theta power towards angry expressions, and decreased right hemispheric alpha power. • Increases in alpha power towards angry than towards neutral expressions.

accuracy was obtained in discriminating the groups by computing a [alpha/theta] ratio, especially when measured in the frontal regions. Moreover, several factors related to different etiologies can explain the clinical symptoms of AD, such as the level of neurodegeneration and the accumulation of the amyloid-beta peptide. Separating groups based on these two variables, Gaubert et al. (62) reported that the most notable effects of neurodegeneration on EEG measures were concentrated in the frontocentral regions. This was marked by a rise in high-frequency oscillations (i.e., higher beta and gamma power), along with a decline in low-frequency oscillations (i.e., lower delta power). In addition, when measuring changes in EEG features after taking amyloid burden into account, the authors reported heterogeneity in participants where the extent of amyloid-beta accumulations can lead to differential spectral power profiles.

Numerous studies have also used EEG to explore brain activity related to emotional processing (67–69). For instance, greater activity in the left prefrontal cortex was found to be associated with approach- related positive emotions, while greater activity in the right prefrontal cortex was associated with withdrawal-related negative emotions (63). In a study by Kisley et al. (64), the researchers examined the late positive potentials (LPP) [i.e., event-related-potentials (ERPs) reflecting enhanced attention to emotional stimuli] in adults ranging from 18 to 81 years old. They found that the LPP amplitudes towards negative images declined linearly with age but remained consistent across ages for positive images. Moreover, Tsolaki et al. (65), reported that healthy older adults demonstrated larger N170 amplitudes than healthy young adults when viewing facial images displaying anger and fear expressions. Despite these prolific findings, few studies have delved into the impairment of facial recognition in elders with dementia using EEG. One recent study reported that AD patients were shown to have lower theta power than healthy controls when perceiving angry facial expressions (66), suggesting the possible implication of EEG for assessing emotional processing in patients with neurocognitive disorders.

Overall, there is an agreement that there is a decrease in EEG activity in cognitive decline, with higher relative spectral power in

the slower frequencies when compared to cognitively unimpaired participants. However, the accessibility and the applicability of EEG sensor devices limit the usage of EEG signals for cognitive skill evaluation, especially for the cognitive decline measures for the aging population. Thus, using findings in spectral power across different frequency bands to validate features from facial mobility could help in identifying which features from the automated facial expression analysis are relevant in the remote assessment of cognitive decline in the elderly population.

3.3 Heart rate and emotion recognition

In recent years, a strong effort has been made to develop contactless technologies to monitor health through physiological measures. Among them, rPPG has been used increasingly in the medical field to assess heart rate (50) and further introduced in emotion analysis (see Table 3 for summary of reviewed articles). Heart rate variability (HRV) and heart rate (HR) as in beats per minute (bpm), while typically measured through electrocardiogram, have successfully been studied using PPG technologies (70, 74). Indeed, it is now believed that HRV can serve as a basis for recognizing emotions, detect stress and overall identify changes in the Autonomic Nervous System (71, 75). According to a systematic review conducted by Cheng et al. (72), patients with dementia or neurocognitive disorders generally exhibit lower resting HRV indices compared to healthy controls. However, after distinguishing between different types of disorders, significant differences in HRV values are observed only in patients with Dementia with Lewy Bodies and MCI. On the contrary, there are no significant differences between patients with AD, Vascular Dementia, and Frontotemporal Dementia and the healthy controls. Furthermore, rPPG has been used in the study of pain and detection of engagement (14, 51). Because physiological, cognitive, and affective events can cause fluctuations in HRV, rPPG can effectively isolate these changes and attribute them to various states (50). Software platforms, such as the FaceReader, have been used in multiple research studies as the heart rate monitoring tools. Great results have been found using this technology, and it remains the most accessible and well-developed rPPG on the market (14). However, accuracy of the physiological monitoring with these software tools remains to be further validated, and their individual usage as a cognitive assessment tool also requires further testing (73).

3.4 Data fusion of heart rate, EEG and AFEA

In an effort to increase accuracy in emotion analysis, some studies have paired rPPG measures with AFEA to establish meaningful correlations between the facial expressions and the physiological measures of emotions (49, 70; see Table 4 for summary of reviewed articles). Interestingly, this pairing allows for both strong (i.e., surprise, fear, joy, etc.) and subtle (stress, contempt, etc.) affective states to be identified. rPPG relies on the discrete changes in heart rate to identify these subtle emotions, while AFEA is successful at differentiating between strong emotions that elicit similar variations in heart rate (77). This fusion of measures ensures that micro-expressions, notorious for escaping AFEAs due to their lack of intensity, are still detectable and accounted for (78).

TABLE 3 Summary of previous research on cognitive decline and heart rate.

Citation	Article type	Sample	Measures included	Stimuli	Main findings
Lu et al. (70)	Empirical	42 healthy adults	Resting electrocardiogram (ECG); Resting earlobe PPG	N/A	• Correlations in the temporal and frequency domains and in nonlinear dynamic analyses between HRV indices derived from PPG and ECG. • PPG can be a practical alternative to ECG for HRV analysis.
Benezeth et al. (71)	Empirical	16 healthy adults	Camera-based rPPG; Contact sensor-based PPG	Video datasets of participants watching videos eliciting fear or anxiety	• High agreement between the HRV analyses derived from the camera data and contact sensor. • Strong correlation between the remote HRV feature and different emotional states.
Cheng et al. (72)	Systematic review and meta-analysis	Dementia patients healthy controls	N/A	N/A	• Lower resting HRV in dementia patients for parasympathetic functions and total variability compared to HCs. • Lower HRV in patients with MCIs and with Dementia with Lewy Bodies compared to HCs. • Lower HRV in patients with Dementia with Lewy Bodies compared to ADs.
Castillo et al. (14)	Empirical	2 elders with dementia 2 healthy elders	rPPG; Manual FACS coding; Video Magnification (VM) algorithm	Video datasets of pain patients	• Correlation between automated FaceReader™ HR estimates and the optimized VM algorithm in baseline and pain conditions. • Correlation between non-verbal automated FaceReader™ pain scores and manual FACS coding. • rPPG can be useful for the automated estimations of HR values and non-verbal pain scores.
Benedetto et al. (73)	Empirical	24 healthy adults	rPPG; ECG	Stress test (i.e., Go/No-Go task)	• Poor accuracy in FaceReader™ rPPG compared to ECG, especially for lower and higher heart rates. • Lack of studies validating consumer devices and more assessment should be conducted.

TABLE 4 Summary of previous research on data fusion of AFEA, rPPG, EEG, and other measures.

Citation	Article type	Sample	Measures included	Stimuli	Main findings
Monkaresi et al. (51)	Empirical	23 healthy adults	Heart Rate (HR); Facial-Feature Based Detection of Engagement	Essay writing	• Accuracy of 75.8% when detecting engagement with pairing facial features and HR in live recordings. • Results were best when data fusion was used, over just facial-feature based detection.
Pham and Wang (76)	Empirical	24 healthy adults	Implicit photoplethysmography (PPG); Facial expression analysis (FEA)	Video advertisements clips	• AttentiveVideo achieved good accuracy (73.59%) on a wide range of emotional measures. • FEA works better for strong emotions (e.g., joy and anger), the PPG channel is more informative for subtle responses or emotions.
McDuff (77)	Doctoral Thesis	N/A	Automated facial expression analysis; Remote measurement of physiology	N/A	• There are clear trends within the physiological responses of individuals and the affect of the content they are watching. • Occurrences of positive valence expressions were predictors of increased preference toward presented stimuli.
Lei et al. (78)	Empirical	Healthy adults	Automated facial expression analysis (iMotions); Galvanic skin response (GSR); Heart Rate (HR)	Emotional videos	• Higher correlation between emotion and GSR compared to emotion and heart rate. • Within a participant, there was no distinct pattern found with the levels of the three parameters measured.
Nagasawa et al. (79)	Empirical	35 health adults	Electroencephalograph (EEG); Automated Facial Expression Analysis; Heart Rate (HR)	FilmStim database	• Noncontact measurement features can be estimated more accurately than EEG extracted features. • Compared to using only facial expressions, combining multiple physiological signals like HR enabled more accurate estimations.
Sun et al. (80)	Empirical	12 healthy adults	Functional near-infrared spectroscopy (fNIRS); Electroencephalograph (EEG); Automated Facial Expression Analysis	Emotional videos	• Results reveal a strong correlation between spontaneous facial affective expressions and the emotional valence. • The affective states were estimated by the fNIRS + EEG brain activity measurements. • Joint utilization of facial expression and wearable neuroimaging for improved emotional analysis.
Koelstra and Patras (81)	Empirical	24 healthy participants	Electroencephalograph (EEG); Automated facial expression analysis	Film clips	• A feature-level fusion approach is demonstrated to improve upon single modality results. • The differences are small and the number of samples too limited to provide a definite answer on the benefits of fusion.

While heart rate monitoring and EEG measures have both individually been paired with automated facial expression analysis establishing correlations, only one recent study investigated employing a multimodal method to increase the accuracy in evaluating emotional states. Nagasawa et al. (79), presented participants with emotion-eliciting videos and obtained their facial recordings as well as EEG signals. Facial recordings were later analyzed to extract physiological responses (i.e., facial expressions, HR, and changes in pupil diameter). After performing an estimation on all data, researchers correlated them with participants' subjective ratings. Results showed a stronger correlation between the estimated arousal signal derived from physiological responses and subjective ratings, compared to those derived from EEG signals, and a similar trend was observed for valence. Therefore, it appears that a multimodal measurement does improve the accuracy of estimating emotions to some extent.

Establishing links between these three measures is imperative in the study of emotions, mostly because they all serve different purposes. If one of these factors can be measured, inferences can be made about the state of the other two. EEG signals can establish the reference value of emotions one is feeling, even if they are not facially expressed (i.e., sadness while smiling). HRV and HR signals are especially indicative of subtle emotions, as seen in prior literature (74, 76). AFEA performs well when detecting strong emotions that are visible through facial movements. Hence, they are all necessary in their own rights in emotion detection. EEG requires extensive equipment and professional guidance to be accurately performed, which is not feasible in the context of remote and contactless emotion analysis, thus only rPPG and AFEA can be used. Considering the established correlations between heart rate variability and brain signals, EEG might not be indispensable in this context. True emotions can be attributed based on heart rate monitoring and therefore replacing EEG in emotion detection. In the case of establishing these correlations with contactless technologies, one would need to conduct a joint study to ensure that past correlations that have been found in EEG signals, heart rate monitoring and facial expression analysis still hold true in contactless technologies (rPPG and AFEA).

In the case of AFEA and EEG specifically, several studies have shown that EEG data can be used to classify different emotion categories processed by participants. Wang et al. (82) reported that the power spectrum was the best EEG analysis method to classify the emotional valence (i.e., positive, or negative) of the stimuli presented. In this study, higher frequency bands (i.e., beta and gamma frequency bands) were shown to have increased robustness at discriminating the valence component of emotions. In addition,

a classification algorithm using the spectral power on different channels was able to classify both the emotional valence and arousal of the emotion processed by participants with a high accuracy (83). These findings indicate that using a relatively basic analysis method of the EEG signal such as spectral power can provide insight into certain components of the emotions being processed by participants, such as arousal and valence. For instance, the combination of EEG features and spontaneous facial expression leads to high accuracy in emotional valence classification (80). This suggests a potential relationship between EEG activity and facial expressions regarding emotional processing, and each of these modalities can offer unique insights. Furthermore, when comparing EEG and facial features on different dimensions of emotional processing, it has been shown that both modalities perform equally at classifying arousal, but that EEG was better at classifying the valence of the emotional stimuli (81). Thus, it appears that facial features can inform about the integrity of emotional processing with an accuracy as good as EEG. This increases the confidence in using automated facial expression analysis to assess emotional processing and as it was discussed in the first section, it is possible to extend this to the assessment of cognitive integrity.

4 Discussion

In the present review, we have highlighted three interconnected cognitive conditions across the elderly population that lack easily accessible, non-invasive detection and progression methods: dementia, MCI, and geriatric depression. More specifically, facial expressions and emotional responses, clear indicators of cognitive decline, have yet to be utilized in the clinical assessment of these conditions. The findings reported here show that there is a link to be made between facial expression features and cognition by assessing emotional processing. We therefore put forth the use of facial expression analysis, augmented by physiological measurements, within the established assessment of these conditions to enhance the accessibility of treatment and progress tracking for the elderly.

As stated in the earlier sections of this review, the current state of the methods used in this clinical area leads to the conclusion that the remote assessment of automated facial expression analysis through the presentation of emotionally charged stimuli with the purpose of assessing cognitive integrity should be further investigated. Given that we can observe changes in muscle tone and activity through passive viewing of such images, the monetary and time cost of cognitive evaluation could be significantly reduced. Although promising, the links between facial expressions during emotional states and cognitive health needs to be validated across various conditions, particularly for the aging population where various levels of cognitive deficits might be present. Hence, the validation of this assessment with the use of EEG analyses will provide increased confidence in the development of robust methods of remote cognitive decline detection.

The potential avenues that stem from these technological developments are not negligible. If facial expression analysis is validated as a viable tool to as an indicator of the progression of cognitive health, the necessary technologies could be implemented within the care centers (i.e.,: a nursing home, a chronic care facility or a residence for seniors) where the elderly are living. The monitoring of their conditions can therefore occur daily via cameras, for example, placed in common living areas and information can be automatically extracted and analyzed by their healthcare provider. This significantly reduces the need for mobility for the elderly to access continuous healthcare. The movement towards automated and in-house health monitoring is already underway, with many products now available to connect individuals to their provider in the comfort of their homes [see Philip et al. (84) for a review of the current technologies for at-home health monitoring for the elderly].

Overall, the combined use of these technologies in emotion recognition provides an increase in accuracy, for both strong and subtle emotions and states. Through such methods, one could potentially obtain true affective states while analyzing the expressed facial movements in order to better understand cognition and emotion processing. These technologies would allow us to move health and medical monitoring into a completely automated phase, in which minimal professional input is needed while profiting the patients. Future work should focus on establishing valid and reliable links between emotional facial expressions and brain activity as well as testing the acceptance of such technologies in the elderly population.

Author contributions

DJ: Conceptualization, Project administration, Supervision, Writing – original draft, Writing – review & editing. LY: Resources, Validation, Writing – review & editing. FM: Conceptualization, Investigation, Project administration, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article.

This research was supported by two graduate fellowships from the NSERC-CREATE network (FM and LY).

Acknowledgments

The authors would like to thank the support for the Aging in Place Challenge Program from the National Research Council of Canada (NRC), as well as Negar Mazloum, Jimmy Hernandez and Linda Pecora for their support and feedback on this review.

Conflict of interest

The 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

References

1. Gaugler J, James B, Johnson T, Reimer J, Solis M, Weuve J, et al. 2022 Alzheimer's disease facts and figures. *Alzheimers & Dementia*. (2022) 18(4):700–89. doi: 10.1002/alz.12638

2. Alzheimer Society of Canada. *Navigating the Path Forward for Dementia in Canada: The Landmark Study Report #1*. Canada: Alzheimer Society of Canada (2022). Available online at: http://alzheimer.ca/en/research/reports-dementia/landmark-study-report-1-path-forward (cited August 13, 2023)

3. Statistics Canada. *2011 Census of Population: Families, Households, Marital status, Structural Type of Dwelling, Collectives*. Ottawa: The Daily (2012).

4. Statistics Canada. *A Portrait of Canada's Growing Population Aged 85 and Older from the 2021 Census*. Ottawa: Statistics Canada (2022).

5. Statistics Canada. *Table 98-10-0045-01 Type of Collective Dwelling, age and Gender for the Population in Collective Dwellings: Canada, Provinces and Territories*. Ottawa: Statistics Canada (2022).

6. Statistics Canada. *Census Profile, 2021 Census of Population: Profile Table*. Ottawa: Statistics Canada (2021).

7. Gustafson L. What is dementia? *Acta Neurol Scand*. (1996) 94:22–4. doi: 10.1111/j.1600-0404.1996.tb00367.x

8. Cerejeira J, Lagarto L, Mukaetova-Ladinska EB. Behavioral and psychological symptoms of dementia. *Front Neurol*. (2012) 3:73. doi: 10.3389/fneur.2012.00073

9. Kunz M, Scharmann S, Hemmeter U, Schepelmann K, Lautenbacher S. The facial expression of pain in patients with dementia. *Pain*. (2007) 133(1-3):221–8. doi: 10.1016/j.pain.2007.09.007

10. Jiang Z, Seyedi S, Haque RU, Pongos AL, Vickers KL, Manzanares CM, et al. Automated analysis of facial emotions in subjects with cognitive impairment. *PLoS One*. (2022) 17(1):e0262527. doi: 10.1371/journal.pone.0262527

11. Seidl U, Lueken U, Thomann PA, Kruse A, Schröder J. Facial expression in Alzheimer's disease: impact of cognitive deficits and neuropsychiatric symptoms. *Am J Alzheimer's Dis Other Demen*. (2012) 27(2):100–6. doi: 10.1177/1533317512440495

12. Burton KW, Kaszniak AW. Emotional experience and facial expression in Alzheimer's disease. *Aging Neuropsychol Cogn*. (2006) 13(3-4):636–51. doi: 10.1080/13825580600735085

13. Henry JD, Rendell PG, Scicluna A, Jackson M, Phillips LH. Emotion experience, expression, and regulation in Alzheimer's disease. *Psychol Aging*. (2009) 24(1):252. doi: 10.1037/a0014001

14. Castillo LI, Browne ME, Hadjistavropoulos T, Prkachin KM, Goubran R. Automated vs. manual pain coding and heart rate estimations based on videos of older adults with and without dementia. *J Rehabil Assist Technol Eng*. (2020) 7:2055668320950196. doi: 10.1177/2055668320950196

15. Bruscoli M, Lovestone S. Is MCI really just early dementia? A systematic review of conversion studies. *Int Psychogeriatr*. (2004) 16(2):129–40. doi: 10.1017/S1041610204000092

16. Bai W, Chen P, Cai H, Zhang Q, Su Z, Cheung T, et al. Worldwide prevalence of mild cognitive impairment among community dwellers aged 50 years and older: a meta-analysis and systematic review of epidemiology studies. *Age Ageing*. (2022) 51(8):afac173. doi: 10.1093/ageing/afac173

17. Roberts RO, Knopman DS, Mielke MM, Cha RH, Pankratz VS, Christianson TJ, et al. Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal. *Neurology*. (2014) 82(4):317–25. doi: 10.1212/wnl.0000000000000055

18. Fei Z, Yang E, Li DD, Butler S, Ijomah W, Zhou H. A survey on computer vision techniques for detecting facial features towards the early diagnosis of mild cognitive impairment in the elderly. *Syst Sci Control Eng*. (2019) 7(1):252–63. doi: 10.1080/21642583.2019.1647577

19. Arevalo-Rodriguez I, Smailagic N, Roqué-Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A, et al. Mini-mental state examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev*. (2021) 7(7):1–68. doi: 10.1002/14651858.CD010783.pub3

20. Chen KH, Lwi SJ, Hua AY, Haase CM, Miller BL, Levenson RW. Increased subjective experience of non-target emotions in patients with frontotemporal dementia and Alzheimer's disease. *Curr Opin Behav Sci*. (2017) 15:77–84. doi: 10.1016/j.cobeha.2017.05.017

21. Pressman PS, Chen KH, Casey J, Sillau S, Chial HJ, Filley CM, et al. Incongruences between facial expression and self-reported emotional reactivity in frontotemporal dementia and related disorders. *J Neuropsychiatry Clin Neurosci*. (2023) 35(2):192–201. doi: 10.1176/appi.neuropsych.21070186

22. Fei Z, Yang E, Yu L, Li X, Zhou H, Zhou W. A novel deep neural network-based emotion analysis system for automatic detection of mild cognitive impairment in the elderly. *Neurocomputing*. (2022) 468:306–16. doi: 10.1016/j.neucom.2021.10.038

23. Hu T, Zhao X, Wu M, Li Z, Luo L, Yang C, et al. Prevalence of depression in older adults: a systematic review and meta-analysis. *Psychiatry Res*. (2022) 311:114511. doi: 10.1016/j.psychres.2022.114511

24. Padayachey U, Ramlall S, Chipps J. Depression in older adults: prevalence and risk factors in a primary health care sample. *S Afr Fam Pract (2004)*. (2017) 59(2):61–6. doi: 10.1080/20786190.2016.1272250

25. Zenebe Y, Akele B, Necho M. Prevalence and determinants of depression among old age: a systematic review and meta-analysis. *Ann Gen Psychiatry*. (2021) 20(1):1–9. doi: 10.1186/s12991-021-00375-x

26. Steffens DC, Skoog I, Norton MC, Hart AD, Tschanz JT, Plassman BL, et al. Prevalence of depression and its treatment in an elderly population: the cache county study. *Arch Gen Psychiatry*. (2000) 57(6):601–7. doi: 10.1001/archpsyc.57.6.601

27. Canadian Institute for Health Information. *Depression among Seniors in Residential Care*. Canada: Canadian Institute for Health Information (2010). Available online at: https://secure.cihi.ca/free_products/ccrs_depression_among_seniors_e.pdf (cited August 13, 2023)

28. Dafsari FS, Jessen F. Depression—an underrecognized target for prevention of dementia in Alzheimer's disease. *Transl Psychiatry*. (2020) 10(1):160. doi: 10.1038/s41398-020-0839-1

29. Ma L. Depression, anxiety, and apathy in mild cognitive impairment: current perspectives. *Front Aging Neurosci*. (2020) 9:95–102. doi: 10.3389/fnagi.2020.00009

30. Muliyala KP, Varghese M. The complex relationship between depression and dementia. *Ann Indian Acad Neurol*. (2010) 13(Suppl2):S69. doi: 10.4103/0972-2327.74248

31. Park SH, Kwak MJ. Performance of the geriatric depression scale-15 with older adults aged over 65 years: an updated review 2000–2019. *Clin Gerontol*. (2021) 44(2):83–96. doi: 10.1080/07317115.2020.1839992

32. Katsikitis M, Pilowsky I. A controlled quantitative study of facial expression in Parkinson's disease and depression. *J Nerv Ment Dis*. (1991) 179(11):683–8. doi: 10.1097/00005053-199111000-00006

33. Gehricke JG, Shapiro D. Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion. *Psychiatry Res*. (2000) 95(2):157–67. doi: 10.1016/S0165-1781(00)00168-2

34. Stolicyn A, Steele JD, Seriès P. Prediction of depression symptoms in individual subjects with face and eye movement tracking. *Psychol Med*. (2022) 52(9):1784–92. doi: 10.1017/s0033291720003608

35. Greden JF, Genero N, Price HL, Feinberg M, Levine S. Facial electromyography in depression: subgroup differences. *Arch Gen Psychiatry*. (1986) 43(3):269–74. doi: 10.1001/archpsyc.1986.01800030087009

36. Sayette MA, Cohn JF, Wertz JM, Perrott MA, Parrott DJ. A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *J Nonverbal Behav*. (2001) 25:167–85. doi: 10.1023/A:1010671109788

37. Cuculo V, D'Amelio A. OpenFACS: an open source FACS-based 3D face animation system. In: *Lecture Notes in Computer Science*. Milan: International Conference on Image and Graphics (2019). p. 232–42. doi: 10.1007/978-3-030-34110-7_20

38. van der Struijk S, Huang H-H, Mirzaei MS, Nishida T. *Facsvatar. Proceedings of the 18th International Conference on Intelligent Virtual Agents* (2018). doi: 10.1145/3267851.3267918

39. Ekman P, Rosenberg EL. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. 2nd ed. Oxford: Oxford Academic (2005). doi: 10.1093/acprof:oso/9780195179644.001.0001

40. Brick TR, Hunter MD, Cohn JF. *Get the FACS fast: automated FACS face analysis benefits from the addition of velocity. 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* IEEE (2009). p. 1–7. doi: 10.1109/acii.2009.5349600

41. Cohn JF, Zlochower AJ, Lien J, Kanade T. Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. In: Ekman P, Rosenberg EL, editors. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. 2nd ed. Oxford: Oxford University Press (2005). p. 371–88. doi: 10.1093/acprof:oso/9780195179644.003.0018

42. Pantic M, Rothkrantz LJ. Automatic analysis of facial expressions: the state of the art. *IEEE Trans Pattern Anal Mach Intell*. (2000) 22(12):1424–45. doi: 10.1109/34.895976

43. Lewinski P. Automated facial coding software outperforms people in recognizing neutral faces as neutral from standardized datasets. *Front Psychol*. (2015) 6:1386. doi: 10.3389/fpsyg.2015.01386

44. Skiendziel T, Rösch AG, Schultheiss OC. Assessing the convergent validity between the automated emotion recognition software noldus FaceReader 7 and facial action coding system scoring. *PLoS One*. (2019) 14(10):e0223905. doi: 10.1371/journal.pone.0223905

45. Dupré D, Krumhuber EG, Küster D, McKeown GJ. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLoS One*. (2020) 15(4):e0231968. doi: 10.1371/journal.pone.0231968

46. Kulke L, Feyerabend D, Schacht A. A comparison of the affectiva iMotions facial expression analysis software with EMG for identifying facial expressions of emotion. *Front Psychol*. (2020) 11:329. doi: 10.3389/fpsyg.2020.00329

47. iMotions. *Facial Expression Analysis: The Complete Pocket Guide*. Copenhagen: iMotions (2020). Available online at: https://imotions.com/blog/facial-expression-analysis/ (cited August 13, 2023)

48. Bandini A, Orlandi S, Escalante HJ, Giovannelli F, Cincotta M, Reyes-Garcia CA, et al. Analysis of facial expressions in Parkinson's disease through video-based automatic methods. *J Neurosci Methods*. (2017) 281:7–20. doi: 10.1016/j.jneumeth.2017.02.006

49. Wang X, Wang Y, Zhou M, Li B, Liu X, Zhu T. Identifying psychological symptoms based on facial movements. *Front Psychiatry*. (2020) 11:607890. doi: 10.3389/fpsyt.2020.607890

50. Sperandeo R, Di Sarno AD, Longobardi T, Iennaco D, Mosca LL, Maldonato NM. Toward a technological oriented assessment in psychology: a proposal for the use of contactless devices for heart rate variability and facial emotion recognition in psychological diagnosis. In: Miglino O, Ponticorvo M, editors. *PSYCHOBIT*. Aachen: CEUR-WS.org (2019). p. 1–7.

51. Monkaresi H, Bosch N, Calvo RA, D'Mello SK. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Trans Affect Comput*. (2016) 8(1):15–28. doi: 10.1109/taffc.2016.2515084

52. Viola P, Jones M. *Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*; IEEE Dec 8, 2001 1. p. I–I. doi: 10.1109/cvpr.2001.990517

53. Loijens L, Krips O. *FaceReader Methodology Note*. Wageningen: Noldus Information Technology Inc (2019). Available online at: https://www.noldus.com/resources/pdf/noldus-white-paper-facereader-methodology.pdf (cited August 13, 2023)

54. Lewinski P, Den Uyl TM, Butler C. Automated facial coding: validation of basic emotions and FACS AUs in FaceReader. *J Neurosci Psychol Econ*. (2014) 7(4):227. doi: 10.1037/npe0000028

55. Algaraawi N, Morris T. Study on aging effect on facial expression recognition. In: Ao SI, Gelman L, Hukins DWL, Hunter A, Korsunsky AM, editors. *Proceedings of the World Congress on Engineering*. London and Hong Kong: Newswood Limited (2016) 1. p. 465–70. Available online at: http://www.iaeng.org/publication/WCE2016/

56. Monk E. *The Monk Skin Tone Scale*. Cambridge, MA: SocArXiv (2023). Available online at: osf.io/preprints/socarxiv/pdf4c

57. Taati B, Zhao S, Ashraf AB, Asgarian A, Browne ME, Prkachin KM, et al. Algorithmic bias in clinical populations—evaluating and improving facial analysis technology in older adults with dementia. *IEEE Access*. (2019) 7:25527–34. doi: 10.1109/access.2019.2900022

58. Buolamwini J, Gebru T. *Gender shades: intersectional accuracy disparities in commercial gender classification. Conference on Fairness, Accountability and Transparency* PMLR (2018). p. 77–91

59. Abrami A, Gunzler S, Kilbane C, Ostrand R, Ho B, Cecchi G. Automated computer vision assessment of hypomimia in Parkinson disease: proof-of-principle pilot study. *J Med Internet Res*. (2021) 23(2):e21037. doi: 10.2196/21037

60. Meghdadi AH, Stevanović Karić M, McConnell M, Rupp G, Richard C, Hamilton J, et al. Resting state EEG biomarkers of cognitive decline associated with Alzheimer's disease and mild cognitive impairment. *PLoS One*. (2021) 16(2):e0244180. doi: 10.1371/journal.pone.0244180

61. Özbek Y, Fide E, Yener GG. Resting-state EEG alpha/theta power ratio discriminates early-onset Alzheimer's disease from healthy controls. *Clin Neurophysiol*. (2021) 132(9):2019–31. doi: 10.1016/j.clinph.2021.05.012

62. Gaubert S, Raimondo F, Houot M, Corsi MC, Naccache L, Diego Sitt J, et al. EEG evidence of compensatory mechanisms in preclinical Alzheimer's disease. *Brain*. (2019) 142(7):2096–112. doi: 10.1093/brain/awz150

63. Palmiero M, Piccardi L. Frontal EEG asymmetry of mood: a mini-review. *Front Behav Neurosci*. (2017) 11:224. doi: 10.3389/fnbeh.2017.00224

64. Kisley MA, Wood S, Burrows CL. Looking at the sunny side of life: age-related change in an event-related potential measure of the negativity bias. *Psychol Sci*. (2007) 18(9):838–43. doi: 10.1111/j.1467-9280.2007.01988.x

65. Tsolaki AC, Kosmidou VE, Kompatsiaris IY, Papadaniil C, Hadjileontiadis L, Tsolaki M. Age-induced differences in brain neural activation elicited by visual emotional stimuli: a high-density EEG study. *Neuroscience*. (2017) 340:268–78. doi: 10.1016/j.neuroscience.2016.10.059

66. Güntekin B, Hanoğlu L, Aktürk T, Fide E, Emek-Savaş DD, Ruşen E, et al. Impairment in recognition of emotional facial expressions in Alzheimer's disease is represented by EEG theta and alpha responses. *Psychophysiology*. (2019) 56(11):e13434. doi: 10.1111/psyp.13434

67. Lazarou I, Adam K, Georgiadis K, Tsolaki A, Nikolopoulos S, (Yiannis) Kompatsiaris I, et al. Can a novel high-density EEG approach disentangle the differences of visual event related potential (N170), elicited by negative facial stimuli, in people with subjective cognitive impairment? *J Alzheimers Dis*. (2018) 65(2):543–75. doi: 10.3233/jad-180223

68. Batty M, Taylor MJ. Early processing of the six basic facial emotional expressions. *Cogn Brain Res*. (2003) 17(3):613–20. doi: 10.1016/s0926-6410(03)00174-5

69. Eimer M, Holmes A. Event-related brain potential correlates of emotional face processing. *Neuropsychologia*. (2007) 45(1):15–31. doi: 10.1016/j.neuropsychologia.2006.04.022

70. Lu G, Yang F, Taylor JA, Stein JF. A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects. *J Med Eng Technol*. (2009) 33(8):634–41. doi: 10.1080/03091900903150998

71. Benezeth Y, Li P, Macwan R, Nakamura K, Gomez R, Yang F. *Remote heart rate variability for emotional state monitoring. 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* IEEE (2018). p. 153–6. doi: 10.1109/bhi.2018.8333392

72. Cheng YC, Huang YC, Huang WL. Heart rate variability in patients with dementia or neurocognitive disorders: a systematic review and meta-analysis. *Aust N Z J Psychiatry*. (2022) 56(1):16–27. doi: 10.1177/0004867420976853

73. Benedetto S, Caldato C, Greenwood DC, Bartoli N, Pensabene V, Actis P. Remote heart rate monitoring-assessment of the facereader rPPg by noldus. *PLoS One*. (2019) 14(11):e0225592. doi: 10.1371/journal.pone.0225592

74. Castaneda D, Esparza A, Ghamari M, Soltanpur C, Nazeran H. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int J Biosens Bioelectron*. (2018) 4(4):195. doi: 10.15406/ijbsbe.2018.04.00125

75. McCraty R. *Science of the Heart: Exploring the Role of the Heart in Human Performance*. Boulder Creek, CA: HeartMath Research Center, Institute of HeartMath (2015). doi: 10.13140/RG.2.1.3873.5128

76. Pham P, Wang J. *Understanding emotional responses to mobile video advertisements via physiological signal sensing and facial expression analysis. Proceedings of the 22nd International Conference on Intelligent User Interfaces* (2017). p. 67–78. doi: 10.1145/3025171.3025186

77. McDuff DJ. *Crowdsourcing affective responses for predicting media effectiveness* (Doctoral dissertation). Massachusetts Institute of Technology.

78. Lei J, Sala J, Jasra SK. Identifying correlation between facial expression and heart rate and skin conductance with iMotions biometric platform. *J Emerg Forensic Sci Res*. (2017) 2(2):53–83.

79. Nagasawa T, Masui K, Doi H, Ogawa-Ochiai K, Tsumura N. Continuous estimation of emotional change using multimodal responses from remotely measured biological information. *Artif Life Robot*. (2022) 27(1):19–28. doi: 10.1007/s10015-022-00734-1

80. Sun Y, Ayaz H, Akansu AN. Multimodal affective state assessment using fNIRS+EEG and spontaneous facial expression. *Brain Sci*. (2020) 10(2):85. doi: 10.3390/brainsci10020085

81. Koelstra S, Patras I. Fusion of facial expressions and EEG for implicit affective tagging. *Image Vis Comput*. (2013) 31(2):164–74. doi: 10.1016/j.imavis.2012.10.002

82. Wang XW, Nie D, Lu BL. Emotional state classification from EEG data using machine learning approach. *Neurocomputing*. (2014) 129:94–106. doi: 10.1016/j.neucom.2013.06.046

83. Liu ZT, Xie Q, Wu M, Cao WH, Li DY, Li SH. Electroencephalogram emotion recognition based on empirical mode decomposition and optimal feature selection. *IEEE Trans Cogn Dev Syst*. (2018) 11(4):517–26. doi: 10.1109/TCDS.2018.2868121

84. Philip NY, Rodrigues JJ, Wang H, Fong SJ, Chen J. Internet of things for in-home health monitoring systems: current advances, challenges and future directions. *IEEE J Sel Areas Commun*. (2021) 39(2):300–10. doi: 10.1109/jsac.2020.3042421

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 23 July 2024
DOI 10.3389/fdgth.2024.1440986

OPEN ACCESS

EDITED BY
Graham Jones,
Tufts Medical Center, United States

REVIEWED BY
Daragh Heitzman,
Texas Neurology, United States
Dale Joachim,
Sonde Health, Inc., United States

*CORRESPONDENCE
Johannes Tröger
johannes.troeger@ki-elements.de

RECEIVED 30 May 2024
ACCEPTED 05 July 2024
PUBLISHED 23 July 2024

CITATION
Tröger J, Dörr F, Schwed L, Linz N, König A, Thies T, Barbe MT, Orozco-Arroyave JR and Rusz J (2024) An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders.
Front. Digit. Health 6:1440986.
doi: 10.3389/fdgth.2024.1440986

An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders

Johannes Tröger[1*], Felix Dörr[1], Louisa Schwed[1], Nicklas Linz[1], Alexandra König[1,2,3], Tabea Thies[4,5], Michael T. Barbe[4], Juan Rafael Orozco-Arroyave[6,7] and Jan Rusz[8]

[1]ki elements GmbH, Saarbrücken, Germany, [2]Cobtek (Cognition-Behaviour-Technology) Lab, University Côte d'azur, Nice, France, [3]Centre de Mémoire de Ressources et de Recherche, Centre Hospitalier Universitaire Nice (CHUN), Nice, France, [4]Department of Neurology, Faculty of Medicine and University Hospital Cologne, Cologne, Germany, [5]IfL Phonetics, Faculty of Arts and Humanities, University of Cologne, Cologne, Germany, [6]GITA Lab, Faculty of Engineering, University of Antioquia, Medellín, Colombia, [7]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, [8]Department of Circuit Theory, Czech Technical University in Prague, Prague, Czechia

Introduction: Dysarthria, a motor speech disorder caused by muscle weakness or paralysis, severely impacts speech intelligibility and quality of life. The condition is prevalent in motor speech disorders such as Parkinson's disease (PD), atypical parkinsonism such as progressive supranuclear palsy (PSP), Huntington's disease (HD), and amyotrophic lateral sclerosis (ALS). Improving intelligibility is not only an outcome that matters to patients but can also play a critical role as an endpoint in clinical research and drug development. This study validates a digital measure for speech intelligibility, the ki: SB-M intelligibility score, across various motor speech disorders and languages following the Digital Medicine Society (DiMe) V3 framework.
Methods: The study used four datasets: healthy controls (HCs) and patients with PD, HD, PSP, and ALS from Czech, Colombian, and German populations. Participants' speech intelligibility was assessed using the ki: SB-M intelligibility score, which is derived from automatic speech recognition (ASR) systems. Verification with inter-ASR reliability and temporal consistency, analytical validation with correlations to gold standard clinical dysarthria scores in each disease, and clinical validation with group comparisons between HCs and patients were performed.
Results: Verification showed good to excellent inter-rater reliability between ASR systems and fair to good consistency. Analytical validation revealed significant correlations between the SB-M intelligibility score and established clinical measures for speech impairments across all patient groups and languages. Clinical validation demonstrated significant differences in intelligibility scores between pathological groups and healthy controls, indicating the measure's discriminative capability.
Discussion: The ki: SB-M intelligibility score is a reliable, valid, and clinically relevant tool for assessing speech intelligibility in motor speech disorders. It holds promise for improving clinical trials through automated, objective, and scalable assessments. Future studies should explore its utility in monitoring disease progression and therapeutic efficacy as well as add data from further dysarthrias to the validation.

KEYWORDS

amyotrophic lateral sclerosis (ALS), Huntington's disease (HD), Parkinson's disease (PD), progressive supranuclear palsy (PSP), speech analysis, intelligibility, digital biomarkers

Introduction

Dysarthria is a motor speech disorder resulting from weakness or paralysis of speech-related muscles (1). It leads to decreased speech intelligibility, frequent communication breakdowns, and a reduced quality of life. Speech intelligibility is reduced in many types of dysarthria, including typical Parkinson's Disease (PD) (2–5), atypical parkinsonism such as progressive supranuclear palsy (PSP) (4, 6, 7), Huntington's disease (HD) (8, 9), amyotrophic lateral sclerosis (ALS) (1, 10), and multiple sclerosis (MS) (11, 12).

Reduced intelligibility of patients' speech often leads to communication difficulties and affects social participation and quality of life in general (13, 14). Hence, communication deficits and perceived intelligibility of their speech represents a major concern for patients with motor speech disorders (15, 16). Speech intelligibility is a construct depending on (a) a speaker (sender) who produces an acoustic signal within, e.g., conversational speech, and (b) a listener (receiver) who receives the signal and interprets it; the success of the interpretation is a direct function of the intelligibility (17) (see also Figure 1). Although a major concern, speech intelligibility is not necessarily dependent on disease severity, duration, or motor phenotype and patients' own perceptions of the severity do not necessarily reflect objective measures (18). Improved intelligibility is often a primary goal of speech therapy, especially for individuals with dysarthria, and can be a valuable endpoint for clinical research and drug development (19).

Accordingly, measuring speech intelligibility is a clinically relevant assessment for monitoring a dysarthric patient's status and tracking the effectiveness of treatments (20). The common method for assessing speech intelligibility is perceptual evaluation by trained personnel—often clinicians. Standard clinical assessments for disorders associated with dysarthria, such as the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) (21), the Unified Huntington's Disease Rating Scale (UHDRS), and the revised amyotrophic lateral sclerosis functional rating scale (ALSFRS-R) (22), are based on clinician-rated questionnaires and assess, among other symptoms, speech intelligibility. However, these assessments require patient and clinician presence and can be subject to observer bias, pointing to a need for more objective automated methods for assessing speech disorders.

As the field of automated speech analysis is growing in clinical research and healthcare applications, there is increasing potential for digital automatic assessments of speech-related symptoms in motor speech disorders (23, 24). Digital dysarthria assessments are better suited for automated patient-administered screening or stratification at low cost to accelerate clinical trials (24–26). Furthermore, a high level of automation can easily scale up outreach to draw unbiased and representative trial populations beyond established clinical sites and hospital networks. In addition, within clinical trials, digital markers deliver objective high-frequency data to guide interventional clinical trial decision-making and make evaluation more efficient (27).

Previous studies have demonstrated how commercially available automatic speech recognition (ASR) systems could be a feasible platform for automatic measures of intelligibility in patients with motor speech disorders (19, 28). As commercial ASR systems are developed majorly on typical—presumably non-dysarthric—speech, the recognition accuracy of such a system should be an inverse model of the intelligibility of the speaker (29–31).

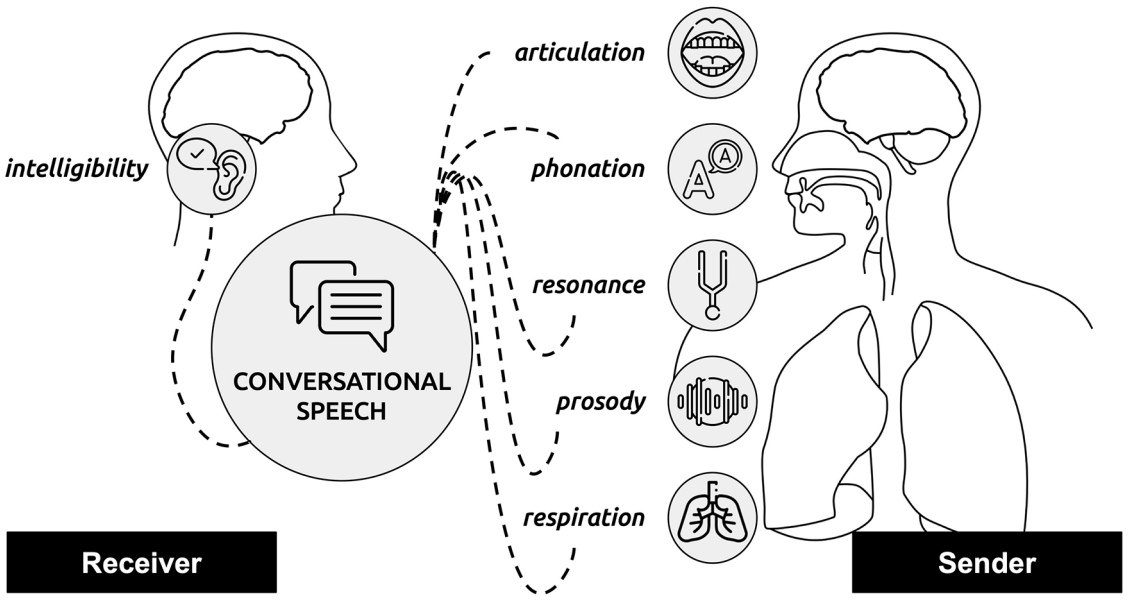


FIGURE 1
Conceptual model of intelligibility; being a receiver/listener-focused measure and being affected by impaired speech subsystems underlying dysarthrias within the sender: articulation, phonation, resonance, prosody, and respiration.

However, although promising results have been published in feasibility studies, there has not been any comprehensive validation work including multiple pathologies and multiple languages and following a systematic validation framework. The Digital Medicine Society (DiMe) V3 framework (verification, analytical validation, and clinical validation) (32–34) defines validation cases that digital measures should comply with to be considered fit-for-purpose for clinical trials and eventually medical devices, such as digital diagnostics. This framework has gained in importance in recent years and can be regarded as an industry standard for digital measures in this field.

In this study, we present a validation following the DiMe V3 framework for a digital measure for intelligibility, the ki: speech biomarker score for motor speech disorders intelligibility (ki: SB-M intelligibility score). We validate the SB-M intelligibility score in individuals with motor speech disorders, including PD, PSP, HD, and ALS, in multiple languages, including German, Czech, and Colombian Spanish, representing the Germanic, Slavic, and Romance language families.

Methods

Data

Four different datasets were used in the analysis: (1) Czech data from $N = 39$ patients with HD (35), $N = 43$ patients with PD (36), $N = 16$ patients with ALS (37), $N = 17$ patients with PSP (6), and $N = 46$ healthy controls (HCs); (2) Colombian data from $N = 50$ HCs and $N = 50$ patients with PD (38); and (3) German data (39) from $N = 98$ patients with PD. For detailed information on the initial cohorts, reading texts, and data collection process, we refer to the initial publications cited; however, for better readability for this manuscript, a short description will be given in the following sections. Compare also Table 1.

Czech data

Participants read an 80-word long paragraph in the respective language, which was phonemically balanced and well-established in clinical research (3). Recordings were conducted in a quiet room with low ambient noise, using a condenser microphone placed approximately 15 cm from the subject's mouth. Each participant had one recording session with the speech-language pathologist, without time limits. Participants were briefed on the speaking tasks and recording process. Each participant provided written informed consent. The collection of the Czech data was approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic (approval number 6/15 Grant GACR VFN).

Colombian data

Participants read 10 sentences of increasing complexity (38). Recordings were collected in a soundproof booth at the Clinica Noel in Medellin, Colombia, using a dynamic omnidirectional microphone and a professional audio card. This study was in compliance with the Helsinki Declaration and was approved by the ethics committee of the Clinica Noel in Medellin, Colombia. Written informed consent was signed by each participant.

German data

Participants read an 80-word long paragraph in the respective language, which was phonemically balanced, well-established, and taken from the German protocol version of the Dysarthria Analyzer (40). Speech data were collected in the Department of Neurology of the University Hospital Cologne in a room with low ambient noise using a condenser microphone headset to keep the mouth-to-microphone distance constant at approximately 7 cm from the mouth. Each participant provided written informed consent. The data collection was approved by the local ethics committee (protocol code: 23-1461-retro).

After the reading task, patients in all three cohorts underwent a range of clinical assessments (different for each study and cohort), of which the following are important for this study: the MDS-UPDRS (21), UHDRS (41), Natural History and Neuroprotection in Parkinson Plus Syndromes—Parkinson Plus Scale (NNIPPS) (42), and ALSFRS-R (22).

Automatic speech recognition and intelligibility score

To calculate the automatic intelligibility scores, we first ran the audios from the reading passage and reading sentences (in Colombian Spanish) through SIGMA the ki: proprietary speech processing library, which—besides other preprocessing and feature extraction steps—also interfaces with commercially available ASR systems; for verification, we selected two different

TABLE 1 Demographic information of the samples and as essential clinical information.

	German	Colombian		Czech				
	PD DE	PD CO	HCs CO	PD CZ	HD CZ	PSP CZ	ALS CZ	HCs CZ
N	98 (32 F)	50 (25 F)	50 (25 F)	43 (19 F)	39 (20 F)	17 (6 F)	16 (11 F)	46 (21 F)
Age (years)	62.7 (±8.23)	61.02(±9.44)	60.98 (±9.46)	63.0 (±9.92)	48.28 (±13.4)	66.76 (±4.8)	60.0 (±10.66)	51.54 (±14.05)
MDS-UPDRS, UHDRS, NNIPPS, ALSFRS-R	37.43 (±10.89)	37.66 (±18.32)	—	20.88 (±10.92)	26.51 (±11.47)	67.12 (±26.7)	35.06 (±6.97)	—
Clinical scale speech items	0.80 (±0.90)	1.34 (±0.82)	—	0.81 (±0.63)	0.81 (±0.46)	1.88 (±0.7)	2.75 (±0.86)	—
ki: SB-M intelligibility score	0.82 (±0.18)	0.73 (±0.18)	0.86 (±0.11)	0.81 (±0.07)	0.67 (±0.17)	0.54 (±0.28)	0.58 (±0.29)	0.85 (±0.04)

CO, Colombian Spanish; CZ, Czech; DE, German.
ALSFRS-R: note that ALSFRS-R has an inverse relationship to disease severity, unlike the other scales where higher scores mean greater severity. Clinical scale speech items: MDS-UPDRS item 3.1, UHDRS dysarthria score, NNIPPS speech item, ALSFRS-R speech item from the bulbar score.

providers: Google Speech API (43) and Amazon Transcribe (44). Based on the transcripts and the target reading texts, we calculated the word error rate (WER, error between the number of target words in the reading text and that in the ASR transcripts) and word accuracy (WA, similar to [28]). From those raw measures, we then derived an automatic proxy for the intelligibility of the speech—the ki: SB-M intelligibility score.

V3 framework

The V3 framework established by the DiMe Society (32) provides a unified evaluation framework for digital measures. V3 includes three distinct phases in sequential order: verification, analytical validation, and clinical validation. For all the three phases, different data have to be collected and statistically analyzed to provide the necessary results.

Verification

Verification entails the systematic evaluation of sample-level sensor outputs against prespecified criteria. The ki: SB-M intelligibility score relies on ASR. Therefore, the most critical part of the sensor output and preprocessing pipeline is the automatic transcription of speech. The ki: SB-M intelligibility score uses a proprietary speech processing pipeline leveraging commercial ASR providers. To verify the performance at this stage, we calculated intraclass correlation coefficients (ICCs) for the WER and SB-M intelligibility score between Google and Amazon ASR. Previous studies and our own work have shown that error rates on a low level, such as phoneme error rate, do not necessarily model losses of perceptual intelligibility (45). We performed verification across the whole data sets except for the German PD data due to a lack of consent from patients.

In addition, we computed ICCs between repeated tests for data sets in which participants performed two repeated reading passages (all CZ data sets). Although tests are executed in quick succession, this can provide first insights into the retest reliability of the measures. Based on the current state of the art in the field, we considered an ICC of 0.40 (fair correlation) acceptable for verification (46).

Analytical validation

Analytical validation evaluates performance to measure a certain concept of interest (similar to construct validity). The ki: SB-M intelligibility score is related to speech impairments resulting in reduced speech intelligibility. For the analytical validation, we compared the ki: SB-M intelligibility score against established clinical anchor measures for speech impairments or dysarthria in the respective populations. Depending on the pathology, these measures differ: PD → MDS-UPDRS → speech item, HD → UHDRS → dysarthria item, PSP → NNIPPS → speech item, and ALS → ALSFRS-R → speech item (please note that in direct comparison with the other clinical scales, the ALSFRS-R has an inverse relationship to disease severity, meaning patients lose points as the disease progresses). For the comparison with the clinical anchors, we computed Spearman's rank correlation coefficient between the ki: SB-M intelligibility score and the respective speech impairment measure.

Clinical validation

Clinical validation evaluates the ability to validly measure clinically meaningful change within an intended scenario, including a specified clinical population. The ki: SB-M intelligibility score is built to measure clinically meaningful change in the intelligibility of speech in dysarthrias. To cover a significant range of dysarthrias, we included clinical validation on the following pathologies: PD, HD, PSP, and ALS.

We performed Kruskal–Wallis test group comparisons in the ki: SB-M intelligibility score between the different diagnostic groups (HC vs. pathology). In addition, we analyzed Spearman's rank correlation between the ki: SB-M intelligibility score and the respective global clinical staging measure: MDS-UPDRS, UHDRS, NNIPPS, and ALSFRS-R.

Results

Verification

For verification of the SB-M intelligibility score, we report reliability between the SB-M intelligibility score based on two different ASR methods and reliability between successive performances of the reading task and calculation of the SB-M intelligibility score.

Inter-rater reliability for ASRs

We compared different ASRs (Google and Amazon) as the basis for the SB-M intelligibility score. For most of the pathological groups, the ICC between both ASR methods showed a good to excellent performance (ICC equal or above 0.30). However, for Colombian PD data, the ICC was only fair and for Czech PD poor; both were still highly significant. The overall HC ICC (across all languages) was also only poor. For details, compare Table 2. WERs showed similar trends to the final intelligibility score, with the following pattern: HCs < PD < HC, PSP = ALS.

Consistency

Consistency over a short period of time (i.e., the same day in the same assessment reading the paragraph twice) was calculated based on repeated paragraph reading in all groups except the Colombian group, which read multiple sentences of increasing difficulty and not one overall homogenous paragraph. The ICCs for consistency were above 0.70, representing a good to excellent agreement. Compare also Table 2.

Analytical validation

For the analytical validation, we compared the ki: SB-M intelligibility score against established clinical anchor measures for speech impairments or dysarthria in the respective

TABLE 2 Agreement between two different ASR methods—Google Speech API and Amazon Transcribe—and the resulting SB-M intelligibility score and raw word error rate.

	HC overall	HC CZ	HC CO	PD CO	PD CZ	HD CZ	PSP CZ	ALS CZ
Google SB-M intelligibility score	0.862 (0.182)	0.853 (0.039)	0.859 (0.200)	0.733 (0.273)	0.810 (0.073)	0.675 (0.173)	0.537 (0.281)	0.590 (0.283)
Amazon SB-M intelligibility score	0.968 (0.088)	0.900 (0.041)	0.980 (0.090)	0.917 (0.177)	0.882 (0.050)	0.775 (0.126)	0.666 (0.28)	0.714 (0.238)
ICC SB-M intelligibility score	0.295 (0.0)	0.180 (0.008)	0.283 (0.0)	0.486 (0.0)	0.290 (0.0)	0.702 (0.0)	0.841 (0.0)	0.869 (0.0)
Google word error rate	0.167 (0.184)	0.238 (0.038)	0.160 (0.2)	0.303 (0.276)	0.288 (0.084)	0.437 (0.154)	0.540 (0.231)	0.479 (0.237)
Amazon word error rate	0.058 (0.113)	0.198 (0.042)	0.032 (0.106)	0.121 (0.202)	0.22 (0.066)	0.372 (0.143)	0.425 (0.228)	0.364 (0.193)
ICC consistency	—	—	—	—	0.75	0.858	0.955	0.982

CO, Colombian Spanish; CZ, Czech.

populations. We found significant correlations between the intelligibility score and the respective dysarthria anchor score for DE PD ($r = -0.46$, $p < 0.01$, $d = 1.03$), CO PD ($r = -0.39$, $p < 0.01$, $d = 0.85$), CZ PD ($r = -0.32$, $p < 0.05$, $d = 0.67$), and CZ HD ($r = -0.37$, $p < 0.05$, $d = 0.80$). Probably owing to the small sample size, statistically we only found a trend in CZ PSP ($r = -0.42$, $p < 0.10$, $d = 0.92$) and CZ ALS ($r = 0.32$, $p = 0.21$, $d = 0.68$), although effect sizes were medium to large. Compare also Figure 2.

Clinical validation

For the group comparisons, we found significant differences, with the ki: SB-M intelligibility score being significantly lower for the respective pathological group for all cohorts: HC CO > PD CO ($H = 17.425$, $p < 0.001$, $\eta^2 = 0.17$), HC CZ > PD CZ ($H = 13.304$, $p < 0.001$, $\eta^2 = 0.14$), CZ HC > CZ HD ($H = 44.437$, $p < 0.001$, $\eta^2 = 0.52$), CZ HC > CZ PSP ($H = 29.696$, $p < 0.001$, $\eta^2 = 0.46$), and CZ HC > CZ ALS ($H = 18.565$, $p < 0.001$, $\eta^2 = 0.29$). For description, please see Table 2, and a graphical overview of the group differences is provided in Figure 3.

Post hoc group comparisons revealed that the intelligibility scores were comparable for the CZ HD, PSP, and ALS groups, and the CZ PD and CO PD groups. However, German PD showed significantly better intelligibility than the other patient groups, actually performing on a par with the other language HC groups.

Discussion

This study aimed to validate the ki: speech biomarker for motor speech disorders intelligibility score (ki: SB-M

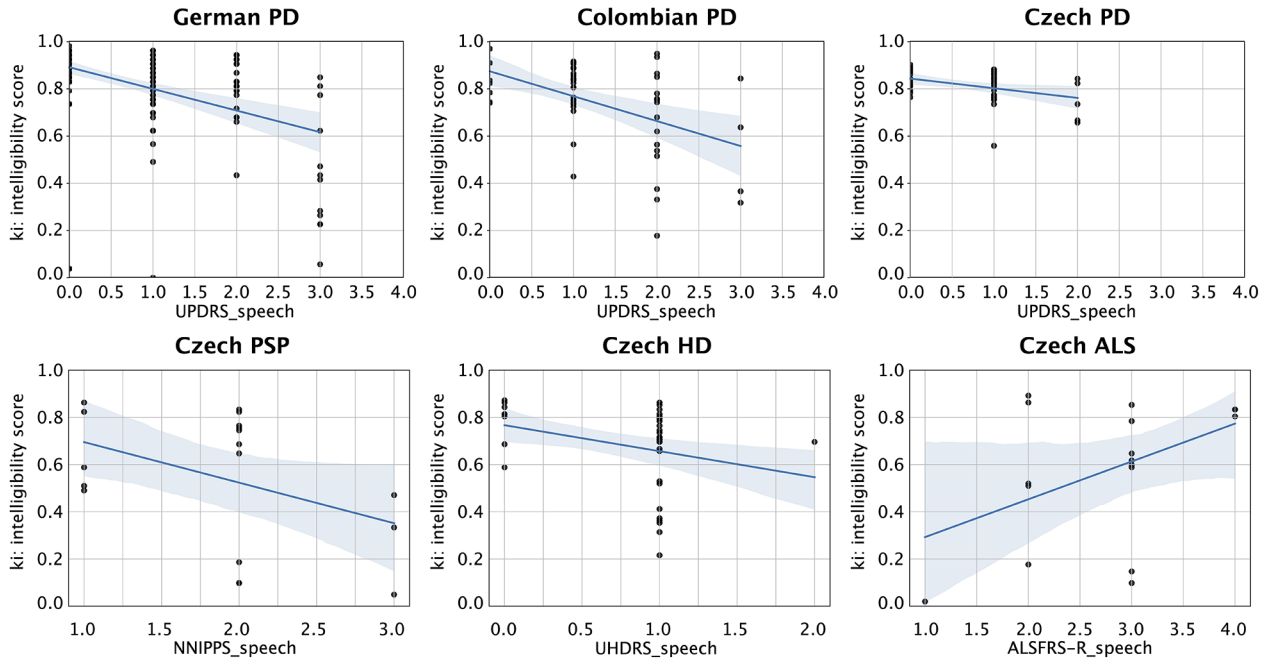


FIGURE 2
Scatter plots for the correlations between the intelligibility score and respective speech dysarthria clinical anchor score. From upper left to lower right: DE PD correlation with the MDS-UPDRS speech item; CO PD correlation with the MDS-UPDRS speech item; CZ PD correlation with the MDS-UPDRS speech item; CZ PSP DE PD correlation with the NNIPPS speech item; CZ HD correlation with the UHDRS dysarthria score; and CZ ALS correlation with the ALSFRS-R speech item (note that ALSFRS-R has an inverse relationship to disease severity, unlike the other scales in which higher scores mean greater severity). DE, German; CO, Colombian Spanish; CZ, Czech.

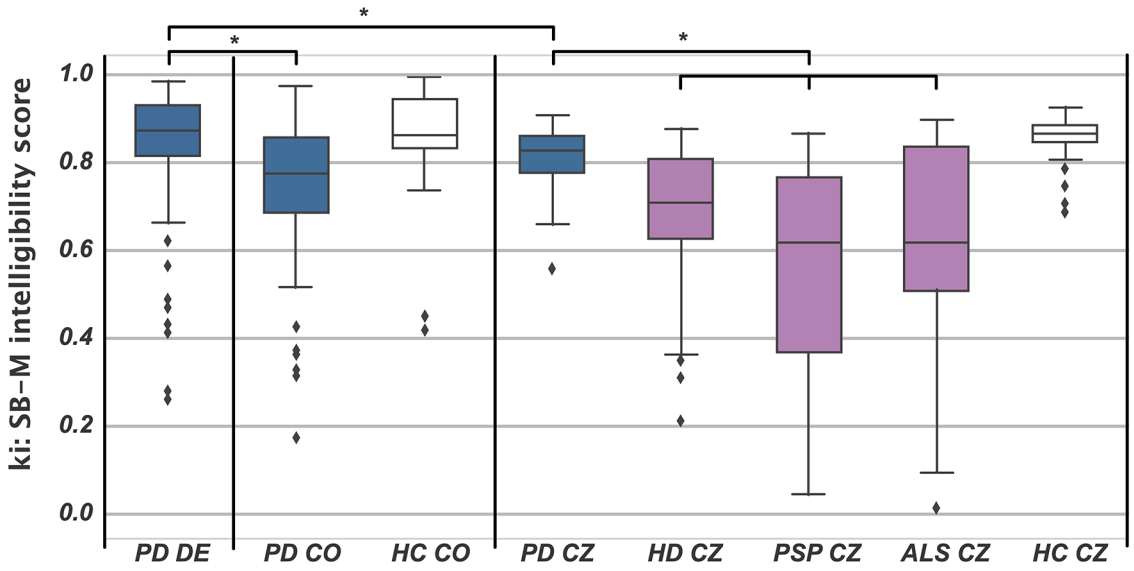


FIGURE 3
Boxplots of the SB-M intelligibility score for all groups. Blue, PD; white, HC; purple, HD, PSP, and ALS. Asterisks denote a significant *post hoc* group comparison. CO, Colombian Spanish; CZ, Czech.

intelligibility score) using the DiMe V3 framework, covering verification, analytical validation, and clinical validation across multiple languages and dysarthria pathologies. Making use of off-the-shelf ASR systems, we took a state-of-the-art approach to automatically measure speech intelligibility in dysarthrias (19, 28, 47). On a conceptual level, we went beyond the aforementioned studies, as we followed the DiME society V3 framework for assessing the readiness of digital measures for clinical research and also included multiple pathologies from the dysarthria spectrum as well as two different ASR systems.

We ran verification on the SB-M intelligibility score, calculating it based on two different automatic speech recognition systems: Google Speech API and Amazon Transcribe. Overall, the ICC indicated good to excellent agreement between the two ASR systems for most pathological groups. However, discrepancies were noted in the Colombian PD and Czech PD data, in which the ICC was only fair to poor, respectively. Poor stability of ASR-based intelligibility measures has been reported previously, especially for typical and mildly impaired severity groups, specifically decreasing their ability to measure changes in the early phases of motor speech disorders (19). The discrepancy might be due to the rather small variance and very good speech recognition, performing almost at an HC level of 0.80, whereas HD, PSP, and ALS have intelligibility scores of 0.70–0.50, with much bigger variances. In these cases, we assume that already small word-level differences inflate discrepancies between ASRs and might cause low ICCs. Especially with the advent of ever-improving ASRs, which also push the needle in dysarthric speech recognition alongside other underrepresented groups, this issue has to be watched closely.

The validity of Google and Amazon ASRs as commercial products naturally extends beyond pathological groups. Both ASR systems have shown high accuracy in recognizing speech from healthy individuals, providing a strong benchmark for comparison (48). However, ensuring robust performance for underrepresented groups remains crucial for the broad applicability and reliability of ASR systems in clinical and everyday settings. On the level of ASR performance in dysarthric speakers, our results compare well with other studies in the field. Gutz et al. (19) found WERs of 10% for mild ALS-related dysarthria to approximately 50% for moderate cases and approximately 80% for severe cases. This is in line with our results for the Czech ALS population, which can be classified as moderately dysarthric based on the ALSFRS-R speech item and shows a 40%–50% WER depending on the ASR system.

Consistency was assessed by comparing the intelligibility scores obtained from repeated paragraph readings. Overall, the ICC values indicated good to excellent consistency. This is an encouraging result but has to be further investigated for repeated measurements of the SB-M intelligibility score assessed longer timeframes apart, such as a couple of days or weeks.

Analytical validation compared the SB-M intelligibility score against established clinical anchor measures for dysarthrias derived from the respective gold standard clinical staging scale. Significant correlations were observed between the SB-M intelligibility score and the respective dysarthria anchor scores for the German PD, Colombian PD, Czech PD, and Czech HD groups. Although specific items are not designed as stand-alone assessments of dysarthria and even less as assessments of intelligibility in principle, we could still demonstrate correlations between the ki: SB-M intelligibility score and those measures. These findings support the SB-M intelligibility score's validity as a measure of perceived speech intelligibility being associated with dysarthria on the speaker side, as confirmed by traditional

clinical assessments such as the MDS-UPDRS, NNIPPS, and ALSFRS-R speech items or the UHDRS dysarthria score. Despite medium to large effect sizes, statistical significance was not achieved for the Czech PSP and Czech ALS groups, likely due to smaller sample sizes. Future studies should aim to include larger cohorts to increase statistical power and provide more robust analytical validation.

Our approach to measuring speech intelligibility differs from other research by using a direct measure based on ASR performance, rather than classifying speech into different states/classes of intelligibility. This research is sometimes carried out using machine learning techniques (49, 50). This line of research frames intelligibility as a classification problem, requiring labeled training data to categorize speech into predefined stages. By contrast, our method leverages the continuous output of ASR systems as a proxy for intelligibility, offering multiple benefits. This continuous measure might provide finer granularity and sensitivity to subtle changes in speech quality over time or between groups. In addition, using an off-the-shelf ASR approach eliminates the need for additional machine learning training, making it more accessible and easier to implement in various clinical and research settings.

One of the major limitations of the analytical validation we performed is that we cannot prove this further by comparing with manual intelligibility ratings by either trained professionals or human raters in general, as has been carried out by Gutz et al. in ALS (19). Future studies should add this piece of analytical validation, leveraging existing methods to rate intelligibility by multiple trained and/or untrained raters (51). In addition, our approach presents, in some respect, a black box approach that directly evaluates dysarthria based on intelligibility as perceived by a somehow non-transparent ASR black box. There is a whole research tradition on using carefully crafted acoustic features to estimate dysarthria and different subsystems, as mentioned in the introduction. Pursuing a hybrid approach that taps into ASR-based intelligibility and traditional acoustic analysis features (e.g., pause rate, articulation rate, pitch instability, or monotonicity) to evaluate patients' dysarthrias would increase the impact of such research and be an important next step.

Clinical validation demonstrated significant differences in SB-M intelligibility scores between healthy controls and pathological groups across all cohorts. This finding underscores the potential of the SB-M intelligibility score as a discriminative tool for identifying and quantifying speech impairments in individuals with motor speech disorders. The consistent pattern of lower intelligibility scores in pathological groups compared with healthy controls across different languages and disorders further supports the robustness and generalizability of the measure. Nevertheless, the experiments presented here still only cover a fraction of the total spectrum of motor-speech-disorder-related dysarthrias or dysarthrias in general. However, our data set of more than 250 patients across four different pathologies and three languages covers a significant amount in this field of research; for rare diseases such as ALS or atypical PD in particular, datasets of that size are rarely reported. In addition, we acknowledge that we did not perform specific testing for cognitive involvement, as the primary aim was to investigate motor speech deviations that are the main contributors to reduced intelligibility. Furthermore, we did not measure the vital capacity of our patients; cohorts such as ALS and PSP may have respiratory impairments that could significantly contribute to reduced intelligibility.

In general, we observed better speech intelligibility in patients with PD than in patients with HD, PSP, or ALS. One reason could be that in the earlier stages of PD, articulation impairment is not as pronounced, allowing for relatively clearer speech. Conversely, HD is characterized by hyperkinetic irregular articulation, and ALS and PSP are associated with hypertonia, leading to imprecise consonant production (52). These speech deficits in HD, ALS, and PSP significantly contribute to reduced intelligibility. These imprecise consonant and uncontrolled (sometimes spastic) irregularities in speech are known to hamper speech intelligibility a lot more than monopitch and monoloudness, which are typically observed in early PD. In addition, the spread in intelligibility scores was a lot greater for HD, PSP, and ALS than for PD, which was also in line with studies on those diseases showing more heterogeneity in their behavioral and speech impairment phenotype.

Between the separate PD groups (DE, CO, and CZ), we observed comparable intelligibility scores in CO and CZ but the German PD group was significantly more intelligible—actually performing on a par with the other language HC groups. This could be related to different recording setups in each study or a general language difference in the underlying ASR performance.

ASR and the measures derived from it exhibit considerable variability when applied to different types of dysarthria (53). Articulatory precision has been identified as the most critical factor influencing speech intelligibility, surpassing the impact of prosody (54).

Finally, another limitation to this study is that we compared intelligibility for audios collected from different studies with different audio recording settings. Although all studies used state-of-the-art microphones for audio recording and professional recording setups—as recommended by recent guidelines (5)—differences in audio recording setups can always play a role in head-to-head comparisons; this is especially the case when comparing our results from CZ directly with CO and DE. Eventually, the accuracy of an automatic speech intelligibility measure is highly dependent on recording conditions. Poor recording environments, such as those with high background noise or subpar microphone quality, can introduce significant bias, leading to artificially low intelligibility ratings. This may result in the erroneous classification of normal speech as dysarthric. Furthermore, different recording devices and handling methods introduce substantial variance, which can confound the measurements and reduce their sensitivity to detect small changes over time or differences between low dysarthria groups. However, one of the most promising scenarios in which to deploy this kind of technology is in at-home environments, where the patient is monitored in everyday life, always using the same device and with similar acoustic conditions. This approach

has shown promising results (55). Future studies in this field should adhere even closer to a standardized recording setup or record with multiple devices—one being a standardized microphone setup next to others.

Conclusion

Overall, this study provides a comprehensive validation of the ki: SB-M intelligibility score for assessing speech intelligibility in motor speech disorders across multiple languages and pathologies. The findings support its reliability, validity, and clinical relevance, highlighting its potential as a standardized tool for clinical and research applications. Automated objective measures of speech intelligibility, such as the SB-M intelligibility score, can increase the efficiency and accuracy of dysarthria assessments, reduce observer bias, and facilitate remote monitoring. This is particularly advantageous for large-scale international clinical trials, in which high-frequency data collection and scalability are critical.

Future efforts should complement validation by investigating the SB-M intelligibility score's ability to monitor disease progression and treatment efficacy. Longitudinal studies assessing changes in the intelligibility score over time and in response to therapeutic interventions could provide valuable insights into the clinical utility of this digital measure.

Data availability statement

The data analyzed in this study are subject to the following licenses/restrictions: the speech data can be accessed from the respective cohort-associated author upon reasonable request. Please navigate through the linked reference for each study in the Methods section. Requests to access these datasets should be directed to rafael.orozco@udea.edu.co, rusz.mz@gmail.com, and tabea.thies@uk-koeln.de.

Ethics statement

The studies involving humans were approved by local ethics committees in Cologne, Germany, Prague, Czech Republic, and Medellín, Colombia. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JT: Writing – original draft, Writing – review & editing. FD: Writing – original draft, Writing – review & editing. LS: Writing – original draft, Writing – review & editing. NL: Writing – original draft, Writing – review & editing. AK: Writing – original draft, Writing – review & editing. TT: Writing – original draft, Writing – review & editing. MB: Writing – original draft, Writing – review & editing. JO-A: Writing – original draft, Writing – review & editing. JR: Writing – original draft, Writing – review & editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article.

JR received grant funding from The Czech Ministry of Health (NW24-04-00211) and National Institute for Neurological Research (Programme EXCELES, ID Project No. LX22NPO5107)—Funded by the European Union—Next Generation EU. The work on the Colombian data was partially funded by CODI at UdeA (PI2023-58010).

Conflict of interest

JT, FD, LS, NL, and AK are employed by the speech biomarker company ki:elements GmbH. JT and NL also hold shares in the speech biomarker company ki:elements GmbH.

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

References

1. Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. *J Speech Hear Res.* (1969) 12(2):246–69. doi: 10.1044/jshr.1202.246

2. Amato F, Borzì L, Olmo G, Artusi CA, Imbalzano G, Lopiano L. Speech impairment in Parkinson's disease: acoustic analysis of unvoiced consonants in Italian native speakers. *IEEE Access.* (2021) 9:166370–81. doi: 10.1109/ACCESS.2021.3135626

3. Rusz J, Cmejla R, Tykalova T, Ruzickova H, Klempir J, Majerova V, et al. Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. *J Acoust Soc Am.* (2013) 134(3):2171–81. doi: 10.1121/1.4816541

4. Rusz J, Tykalová T, Salerno G, Bancone S, Scarpelli J, Pellecchia MT. Distinctive speech signature in cerebellar and parkinsonian subtypes of multiple system atrophy. *J Neurol.* (2019) 266(6):1394–404. doi: 10.1007/s00415-019-09271-7

5. Rusz J, Tykalova T, Ramig LO, Tripoliti E. Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. *Mov Disord.* (2021) 36(4):803–14. doi: 10.1002/mds.28465

6. Daoudi K, Das B, Tykalova T, Klempir J, Rusz J. Speech acoustic indices for differential diagnosis between Parkinson's disease, multiple system atrophy and progressive supranuclear palsy. *NPJ Parkinsons Dis*. (2022) 8(1):1–13. doi: 10.1038/s41531-021-00272-w

7. Kim Y, Kent RD, Kent JF, Duffy JR. Perceptual and acoustic features of dysarthria in multiple system atrophy. *J Med Speech-Lang Pathol*. (2010) 18(4):66–71.

8. Diehl SK, Mefferd AS, Lin YC, Sellers J, McDonell KE, De Riesthal M, et al. Motor speech patterns in Huntington disease. *Neurology*. (2019) 93(22):E2042–52. doi: 10.1212/WNL.0000000000008541

9. Kouba T, Frank W, Tykalova T, Mühlbäck A, Klempíř J, Lindenberg KS, et al. Speech biomarkers in Huntington disease: a cross-sectional study in pre-symptomatic, prodromal and early manifest stages. *Eur J Neurol*. (2023) 1262–71. doi: 10.1111/ene.15726

10. Tomik B, Guiloff RJ. Dysarthria in amyotrophic lateral sclerosis: a review. *Amyotroph Lateral Scler*. (2010) 11(1–2):4–15. doi: 10.3109/17482960802379004

11. Darley FL, Brown JR, Goldstein NP. Dysarthria in multiple sclerosis. *J Speech Hear Res*. (1972) 15(2):229–45. doi: 10.1044/jshr.1502.229

12. Rusz J, Benova B, Ruzickova H, Novotny M, Tykalova T, Hlavnicka J, et al. Characteristics of motor speech phenotypes in multiple sclerosis. *Mult Scler Relat Disord*. (2018) 19:62–9. doi: 10.1016/j.msard.2017.11.007

13. Van Uem JMT, Marinus J, Canning C, Van Lummel R, Dodel R, Liepelt-Scarfone I, et al. Health-related quality of life in patients with Parkinson's disease—a systematic review based on the ICF model. *Neurosci Biobehav Rev*. (2016) 61:26–34. doi: 10.1016/j.neubiorev.2015.11.014

14. Chu SY, Tan CL. Subjective self-rated speech intelligibility and quality of life in patients with Parkinson's disease in a Malaysian sample. *Open Public Health J*. (2018) 11(1):485–93. doi: 10.2174/1874944501811010485

15. McAuliffe MJ, Baylor CR, Yorkston KM. Variables associated with communicative participation in Parkinson's disease and its relationship to measures of health-related quality-of-life. *Int J Speech Lang Pathol*. (2017) 19(4):407–17. doi: 10.1080/17549507.2016.1193900

16. Schrag A, Jahanshahi M, Quinn N. How does Parkinson's disease affect quality of life? A comparison with quality of life in the general population. *Mov Disord*. (2000) 15(6):1112–8. doi: 10.1002/1531-8257(200011)15:6<1112::AID-MDS1008>3.0.CO;2-A

17. Yorkston KM, Yorkston KM. *Management of motor speech disorders in children and adults*. 2nd ed. Austin, TX: Pro-Ed (1999). p. 618.

18. Miller N, Allcock L, Jones D, Noble E, Hildreth AJ, Burn DJ. Prevalence and pattern of perceived intelligibility changes in Parkinson's disease. *J Neurol Neurosurg Amp Psychiatry*. (2007) 78(11):1188–90. doi: 10.1136/jnnp.2006.110171

19. Gutz SE, Stipancic KL, Yunusova Y, Berry JD, Green JR. Validity of off-the-shelf automatic speech recognition for assessing speech intelligibility and speech severity in speakers with amyotrophic lateral sclerosis. *J Speech Lang Hear Res*. (2022) 65 (6):2128–43. doi: 10.1044/2022_JSLHR-21-00589

20. Kent RD, Weismer G, Kent JF, Rosenbek JC. Toward phonetic intelligibility testing in dysarthria. *J Speech Hear Disord*. (1989) 54(4):482–99. doi: 10.1044/jshd.5404.482

21. 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: MDS-UPDRS: clinimetric assessment. *Mov Disord*. (2008) 23(15):2129–70. doi: 10.1002/mds.22340

22. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *J Neurol Sci*. (1999) 169(1–2):13–21. doi: 10.1016/S0022-510X(99)00210-5

23. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. *Digit Biomark*. (2021) 5 (1):78–88. doi: 10.1159/000515346

24. Espay AJ, Hausdorff JM, Sánchez-Ferro Á, Klucken J, Merola A, Bonato P, et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. *Mov Disord*. (2019) 34(5):657–63. doi: 10.1002/mds.27671

25. Salmon K, Genge A. "Clinical trials in ALS—current challenges and strategies for future directions". In: Shaw CA, Morrice JR, editors. *Spectrums of Amyotrophic Lateral Sclerosis*. 1st ed. Hoboken, NJ: John Wiley & Sons, Inc. (2021). p. 161–80.

26. Bowden M, Beswick E, Tam J, Perry D, Smith A, Newton J, et al. A systematic review and narrative analysis of digital speech biomarkers in motor neuron disease. *NPJ Digit Med*. (2023) 6(1):228. doi: 10.1038/s41746-023-00959-9

27. Dorsey ER, Papapetropoulos S, Xiong M, Kieburtz K. The first frontier: digital biomarkers for neurodegenerative disorders. *Digit Biomark*. (2017) 1(1):6–13. doi: 10.1159/000477383

28. Vásquez-Correa JC, Orozco-Arroyave JR, Bocklet T, Nöth E. Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *J Commun Disord*. (2018) 76:21–36. doi: 10.1016/j.jcomdis.2018.08.002

29. De Russis L, Corno F. On the impact of dysarthric speech on contemporary ASR cloud platforms. *J Reliab Intell Environ*. (2019) 5(3):163–72. doi: 10.1007/s40860-019-00085-y

30. Mustafa MB, Rosdi F, Salim SS, Mughal MU. Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. *Expert Syst Appl*. (2015) 42(8):3924–32. doi: 10.1016/j.eswa.2015.01.033

31. Keshet J. Automatic speech recognition: a primer for speech-language pathology researchers. *Int J Speech Lang Pathol*. (2018) 20(6):599–609. doi: 10.1080/17549507.2018.1510033

32. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). *NPJ Digit Med*. (2020) 3(1):1–15. doi: 10.1038/s41746-020-0260-4

33. Coravos A, Doerr M, Goldsack J, Manta C, Shervey M, Woods B, et al. Modernizing and designing evaluation frameworks for connected sensor technologies in medicine. *NPJ Digit Med*. (2020) 3(1):37. doi: 10.1038/s41746-020-0237-3

34. Goldsack JC, Dowling AV, Samuelson D, Patrick-Lake B, Clay I. Evaluation, acceptance, and qualification of digital measures: from proof of concept to endpoint. *Digit Biomark*. (2021) 5(1):53–64. doi: 10.1159/000514730

35. Rusz J, Klempíř J, Tykalová T, Baborová E, Čmejla R, Růžička E, et al. Characteristics and occurrence of speech impairment in Huntington's disease: possible influence of antipsychotic medication. *J Neural Transm*. (2014) 121 (12):1529–39. doi: 10.1007/s00702-014-1229-8

36. Hlavnička J, Čmejla R, Tykalová T, Šonka K, Růžička E, Rusz J. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Sci Rep*. (2017) 7(1):12. doi: 10.1038/s41598-017-00047-5

37. Novotny M, Melechovsky J, Rozenstoks K, Tykalova T, Kryze P, Kanok M, et al. Comparison of automated acoustic methods for oral diadochokinesis assessment in amyotrophic lateral sclerosis. *J Speech Lang Hear Res*. (2020) 63(10):3453–60. doi: 10.1044/2020_JSLHR-20-00109

38. Orozco-Arroyave JR, Arias-Londoño JD, Vargas-Bonilla JF, González-Rátiva MC, Nöth E. *New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease*. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); 26–31 May 2014; Reykjavik, Iceland*. Paris: European Language Resources Association (ELRA) (2014). p. 342–7.

39. Thies T, Mücke D, Lowit A, Kalbe E, Steffen J, Barbe MT. Prominence marking in parkinsonian speech and its correlation with motor performance and cognitive abilities. *Neuropsychologia*. (2020) 137:107306. doi: 10.1016/j.neuropsychologia.2019.107306

40. Hlavnička J. The dysarthria analyzer. Available online at: https://www.dysan.cz/ (accessed June 25, 2024).

41. Kieburtz K, Penney JB, Como P, Ranen N, Shoulson I, Feigin A, et al. Unified Huntington's disease rating scale: reliability and consistency. *Mov Disord*. (1996) 11 (2):136–42. doi: 10.1002/mds.870110204

42. Payan CAM, Viallet F, Landwehrmeyer BG, Bonnet AM, Borg M, Durif F, et al. Disease severity and progression in progressive supranuclear palsy and multiple system atrophy: validation of the NNIPPS—Parkinson plus scale. *PLoS One*. (2011) 6(8):e22293. doi: 10.1371/journal.pone.0022293

43. Google LLC. Google speech-to-text V2.25.1 (2023). Available online at: https://cloud.google.com/speech-to-text/ (Accessed July 29, 2022).

44. Amazon Web Services Inc. Amazon Transcribe V1.20.15 (2023). Available online at: https://aws.amazon.com/pm/transcribe/ (Accessed April 11, 2024).

45. Van Nuffelen G, Middag C, De Bodt M, Martens J. Speech technology-based assessment of phoneme intelligibility in dysarthria. *Int J Lang Commun Disord*. (2009) 44(5):716–30. doi: 10.1080/13682820802342062

46. Ratitch B, Trigg A, Majumder M, Vlajnic V, Rethemeier N, Nkulikiyinka R. Clinical validation of novel digital measures: statistical methods for reliability evaluation. *Digit Biomark*. (2023) 7:74–91. doi: 10.1159/000531054

47. Dimauro G, Di Nicola V, Bevilacqua V, Caivano D, Girardi F. Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system. *IEEE Access*. (2017) 5:22199–208. doi: 10.1109/ACCESS.2017.2762475

48. Zhang Y, Han W, Qin J, Wang Y, Bapna A, Chen Z, et al. Google USM: scaling automatic speech recognition beyond 100 languages. arXiv [preprint]. (2023). Available online at: https://arxiv.org/abs/2303.01037 (accessed July 1, 2024).

49. Kim J, Kumar N, Tsiartas A, Li M, Narayanan SS. Automatic intelligibility classification of sentence-level pathological speech. *Comput Speech Lang*. (2015) 29 (1):132–44. doi: 10.1016/j.csl.2014.02.001

50. Huang A, Hall K, Watson C, Shahamiri SR. *A review of automated intelligibility assessment for dysarthric speakers*. In: *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD); 13–15 Oct 2021; Bucharest, Romania*. New York, NY: IEEE (2021). p. 19–24.

51. Stipancic KL, Tjaden K, Wilding G. Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *J Speech Lang Hear Res*. (2016) 59(2):230–8. doi: 10.1044/2015_JSLHR-S-15-0271

52. Campi M, Peters GW, Toczydlowska D. Ataxic speech disorders and Parkinson's disease diagnostics via stochastic embedding of empirical mode decomposition. *PLoS One*. (2023) 18(4):e0284667. doi: 10.1371/journal.pone.0284667

53. Rowe HP, Gutz SE, Maffei MF, Tomanek K, Green JR. Characterizing dysarthria diversity for automatic speech recognition: a tutorial from the clinical perspective. *Front Comput Sci*. (2022) 4:770210. doi: 10.3389/fcomp.2022.770210

54. Tu M, Wisler A, Berisha V, Liss JM. The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance. *J Acoust Soc Am*. (2016) 140(5):EL416–22. doi: 10.1121/1.4967208

55. Arias-Vergara T, Vásquez-Correa JC, Orozco-Arroyave JR, Nöth E. Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. *Speech Commun*. (2018) 101:11–25. doi: 10.1016/j.specom.2018.05.007

frontiers | Frontiers in Digital Health

TYPE Correction
PUBLISHED 13 November 2024
DOI 10.3389/fdgth.2024.1488178

OPEN ACCESS

APPROVED BY
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

*CORRESPONDENCE
Johannes Tröger
johannes.troeger@ki-elements.de

RECEIVED 29 August 2024
ACCEPTED 31 October 2024
PUBLISHED 13 November 2024

CITATION
Tröger J, Dörr F, Schwed L, Linz N, König A, Thies T, Barbe MT, Orozco-Arroyave JR and Rusz J (2024) Corrigendum: An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders.
Front. Digit. Health 6:1488178.
doi: 10.3389/fdgth.2024.1488178

Corrigendum: An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders

Johannes Tröger[1*], Felix Dörr[1], Louisa Schwed[1], Nicklas Linz[1], Alexandra König[1,2,3], Tabea Thies[4,5], Michael T. Barbe[4], Juan Rafael Orozco-Arroyave[6,7] and Jan Rusz[8]

[1]ki elements GmbH, Saarbrücken, Germany, [2]Cobtek (Cognition-Behaviour-Technology) Lab, University Côte d'azur, Nice, France, [3]Centre de Mémoire de Ressources et de Recherche, Centre Hospitalier Universitaire Nice (CHUN), Nice, France, [4]Department of Neurology, Faculty of Medicine and University Hospital Cologne, Cologne, Germany, [5]IfL Phonetics, Faculty of Arts and Humanities, University of Cologne, Cologne, Germany, [6]GITA Lab, Faculty of Engineering, University of Antioquia, Medellín, Colombia, [7]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, [8]Department of Circuit Theory, Czech Technical University in Prague, Prague, Czechia

KEYWORDS

amyotrophic lateral sclerosis (ALS), Huntington's disease (HD), Parkinson's disease (PD), progressive supranuclear palsy (PSP), speech analysis, intelligibility, digital biomarkers

A Corrigendum on
An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders
By Tröger J, Dörr F, Schwed L, Linz N, König A, Thies T, Barbe MT, Orozco-Arroyave JR and Rusz J (2024). *Front Digit Health* 6: 1440986. doi: 10.3389/fdgth.2024.1440986

In the published article, there was an error in the author list, and author Michael T. Barbe was erroneously excluded. The corrected author list appears below.

Johannes Tröger[1*], Felix Dörr[1], Louisa Schwed[1], Nicklas Linz[1], Alexandra König[1,2,3], Tabea Thies[4,5], Michael T. Barbe[4], Juan Rafael Orozco-Arroyave[6,7] and Jan Rusz[8]

Due to this error the **Author contributions** section has also been updated and the corrected statement appears below.

Author contributions

JT: Writing – original draft, Writing – review & editing. FD: Writing – original draft, Writing – review & editing. LS: Writing – original draft, Writing – review & editing. NL: Writing – original draft, Writing – review & editing. AK: Writing – original draft, Writing – review & editing. TT: Writing – original draft, Writing – review & editing. MB: Writing – original draft, Writing – review & editing. JO-A: Writing – original draft, Writing – review & editing. JR: Writing – original draft, Writing – review & editing.

The authors apologize for this error and states that this does not change the scientific conclusions of the article in any way. The original article has been updated.

Publisher's note

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 26 August 2024
DOI 10.3389/fdgth.2024.1435693

OPEN ACCESS

EDITED BY
Pradeep Nair,
Central University of Himachal Pradesh, India

REVIEWED BY
Jessica De Souza,
University of California, San Diego, United States
Kristin Hannesdottir,
Novartis, United States

*CORRESPONDENCE
T. Kangarloo
✉ tairmae.kangarloo@takeda.com

RECEIVED 20 May 2024
ACCEPTED 02 August 2024
PUBLISHED 26 August 2024

CITATION
Kangarloo T, Latzman RD, Adams JL, Dorsey R, Kostrzebski M, Severson J, Anderson D, Horak F, Stephenson D and Cosman J (2024) Acceptability of digital health technologies in early Parkinson's disease: lessons from WATCH-PD.
Front. Digit. Health 6:1435693.
doi: 10.3389/fdgth.2024.1435693

Acceptability of digital health technologies in early Parkinson's disease: lessons from WATCH-PD

T. Kangarloo[1]*, R. D. Latzman[1], J. L. Adams[2,3], R. Dorsey[2,3], M. Kostrzebski[2,3], J. Severson[4], D. Anderson[4], F. Horak[5], D. Stephenson[6] and J. Cosman[7]

[1]Takeda Pharmaceuticals, Cambridge, MA, United States, [2]Center for Health+Technology, University of Rochester, Rochester, NY, United States, [3]Department of Neurology, University of Rochester, Rochester, NY, United States, [4]Clinical Ink, Horsham, PA, United States, [5]Balance Disorders Laboratory, Department of Neurology, Oregon Health and Science University, Portland, OR, United States, [6]Critical Path Institute, Tuscan, AZ, United States, [7]Abbvie, North Chicago, IL, United States

Introduction: Digital health technologies (DHTs) have the potential to alleviate challenges experienced in clinical trials through more objective, naturalistic, and frequent assessments of functioning. However, implementation of DHTs come with their own challenges, including acceptability and ease of use for study participants. In addition to acceptability, it is also important to understand device proficiency in the general population and within patient populations who may be asked to use DHTs for extended periods of time. We thus aimed to provide an overview of participant feedback on acceptability of DHTs, including body-worn sensors used in the clinic and a mobile application used at-home, used throughout the duration of the Wearable Assessments in the Clinic and at Home in Parkinson's Disease (WATCH-PD) study, an observational, longitudinal study looking at disease progression in early Parkinson's Disease (PD).
Methods: 82 participants with PD and 50 control participants were enrolled at 17 sites throughout the United States and followed for 12 months. We assessed participants' general device proficiency at baseline, using the Mobile Device Proficiency Questionnaire (MDPQ). The mean MDPQ score at Baseline did not significantly differ between PD patients and healthy controls (20.6 [2.91] vs 21.5 [2.94], p = .10).
Results: Questionnaire results demonstrated that participants had generally positive views on the comfort and use of the digital technologies throughout the duration of the study, regardless of group.
Discussion: This is the first study to evaluate patient feedback and impressions of using technology in a longitudinal observational study in early Parkinson's Disease. Results demonstrate device proficiency and acceptability of various DHTs in people with Parkinson's does not differ from that of neurologically healthy older adults, and, overall, participants had a favorable view of the DHTs deployed in the WATCH-PD study.

KEYWORDS

digital tool, patient feedback, Parkinson, wearability, wearable sensors

1 Introduction

Advances in digital technologies, such as mobile phones and wearables, are now ubiquitous and have changed how we interact with others and the world around us. For example, a 2020 poll showed that 90% of Americans own a smartphone and 21% own a smartwatch or fitness tracker (1, 2). Beyond giving us the capabilities to post

pictures, play games, or track our workouts, these technologies have become particularly valuable in the health and research sectors (3). In clinical trials, for example, as opposed to traditional assessments, which are subjective and performed infrequently, digital tools have the potential to provide a more holistic view of disease symptoms (4–6), progression (7–9), and response to treatment (5). Furthermore, using digital tools in fully decentralized or hybrid clinical trials can reduce or fully eliminate site visits, a documented barrier to clinical trial participation due to patient and caregiver burden (10).

Although using digital health technologies (DHTs) may alleviate some of the challenges faced in clinical trials, they often come with their own challenges resulting in lower rates of adaptation, particularly among older individuals. There is a false assumption of device proficiency in the general population, especially when working with a population of older adults, who require greater assistance in relation to digital technologies than younger populations (11). For instance, a nonexperimental study design exploring attitudes about technology in older adults found that older adults were willing to use technology but had negative outlooks associated with technology creating inconveniences and unhelpful features, thus making it harder to use and navigate (12). Other factors that have contributed to low technology adaptation in older adults include poor technology designs that don't consider the perceptual and cognitive abilities of older adults, and poor training on use of the technology (13).

One disease consisting primarily of older adults where the use of DHTs has been especially relevant in clinical trial measurement is Parkinson's Disease (PD). PD, the second most prevalent and fastest growing movement disorder in the world, affects about 1% of adults 60 years and older (14). The cardinal features of the disease are motor impairments such as tremor, rigidity, and bradykinesia, however, the clinical features extend beyond that as patients typically bring to light the cognitive and mood impairments caused by the disease (15, 16). The current gold standard for assessing progression in PD, the Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) (17), has limitations which pose challenges for clinical trials. For instance, to properly power a phase II clinical trial to see a change in the MDS-UPDRS studies must have large sample sizes and long study durations (18, 19). The frequency in which participants need to come into the clinic in traditional clinical trials can also be a hurdle as clinical trials are typically run in large, academic hospitals researchers are only capturing participants that live in metropolitan areas or have the means to travel to study sites (20). Using digital technologies in clinical trials can not only give us better, more sensitive, measures of disease progression but can also help us reach a wider range of participants by reducing the number of clinic visits or potentially shifting towards totally remote clinical trials.

One method to assess comfort with technology in older adults is the Mobile Device Proficiency Questionnaire (MDPQ). The MDPQ includes items related to comfort using devices, such as tablets and smartphones, and has been found to be a highly reliable measure of mobile device proficiency in older adults (21). The MDPQ could serve as a tool to identify participants who may need more training in using digital technologies in clinical trials. Additionally, researchers can evaluate patients' first-hand experiences using DHTs by harnessing the voice of the individuals participating in research studies and clinical trials. Acquiring patient feedback early and often, through panels, interviews, and questionnaires, can provide insights related to the acceptability of these technologies and help inform future study design.

1.1 Current study

The Wearable Assessments in the Clinic and at Home in Parkinson's Disease [WATCH-PD (4);] study was a one-year, observational study exploring disease progression using DHTs in early Parkinson's Disease. Perceptions of the DHTs used in the WATCH-PD study were captured from participants throughout the study. In this paper we aim to give an overview of participant feedback with the goal of providing a better understanding of the feasibility and burden of using these technologies during participation in longitudinal clinical trials. Specifically, we aimed to report if there are differences between people with PD and control participants in (1) device proficiency at baseline as measured by the MDPQ and (2) overall impressions of using digital technologies during participation in a 12-month longitudinal study.

2 Methods

2.1 Trial design

The Wearable Assessments in the Clinic and at Home in PD is a prospective, longitudinal, multisite natural history study in people with early, untreated PD (<2 yr since diagnosis) and neurologically healthy matched controls. 82 participants with PD and 50 control participants were enrolled at 17 sites throughout the United States and followed for 12 months. Participants completed regular clinic visits in addition to completing self-administered assessments of motor and non-motor function outside of the clinic using a mobile application twice monthly. A brief description is provided below. For a fuller description, please see Adams et al. (4).

2.2 Participants

Participants were recruited from clinics, study interest registries, and social media. We aimed to evaluate a population similar to the Parkinson's Progression Markers Initiative (PPMI) (22). Thus, at enrollment, PD patients were required to be aged 30 or older, within 2 years of diagnosis, untreated with symptomatic medications [including levodopa, dopamine agonists, Monoamine oxidase-B (MAO-B) inhibitors, amantadine, anticholinergics] and not expected to require medication for at least 6 months at baseline, a modified Hoehn and Yahr $\leq$2, and at least two of the following symptoms: resting tremor, bradykinesia, or rigidity (must have either resting tremor or bradykinesia as one of two symptoms); OR either asymmetric resting tremor or asymmetric bradykinesia.

Control participants were required to be aged 30 or older at the time of enrollment, with no diagnosis of a significant Central Nervous System (CNS) disease (other than PD), history of repeated head injury, history of epilepsy or seizure disorder other than febrile seizures as a child, or history of a brain magnetic resonance imaging (MRI) scan indicative of clinically significant abnormality. For both PD patients and controls, a Montreal Cognitive Assessment (MoCA) score < 24 was considered exclusionary.

2.3 Study assessments

Each participant completed clinic visits at Screening/Baseline, 1, 3, 6, 9, 12 months. Clinic visits consisted of three core components: (1) a comprehensive battery of clinician and patient-reported outcomes measuring both motor and non-motor function, (2) a set of motor assessments completed while wearing inertial sensors distributed across the body, and (3) completion of a custom-developed, self-administered mobile phone battery designed to measure aspects of motor and non-motor function. In addition to in-clinic assessments, participants were asked to wear a smartwatch on the wrist of their most affected side for 7 days following each clinic visit and were asked to complete the same mobile battery they completed during clinic visits every two weeks for the duration of the study. Due to COVID-19 a subset of individuals did some of the in-clinic assessments remotely and not all data were available.

2.4 Instrumented motor assessments

At each clinic visit, participants were instrumented with a set of six Opal sensors (OPAL system, APDM, Inc., Portland, OR, United States) placed on each wrist, around each foot, and one sensor each positioned on the sternum and the lumbar area (Figure 1). The Opal sensors contain 3-axis, accelerometers, gyroscopes, and magnetometers, and were used to capture raw kinematic data during the performance of the MDS-UPDRS Part III motor examination, as well as a 5× sit-to-stand task, a 30 s standing balance task (eyes open), a two-minute walking task and a two-minute walking task under cognitive load (serial sevens).

2.5 Mobile assessment battery

As noted above, participants were provided with a provisioned smartphone and smartwatch and completed a custom-designed

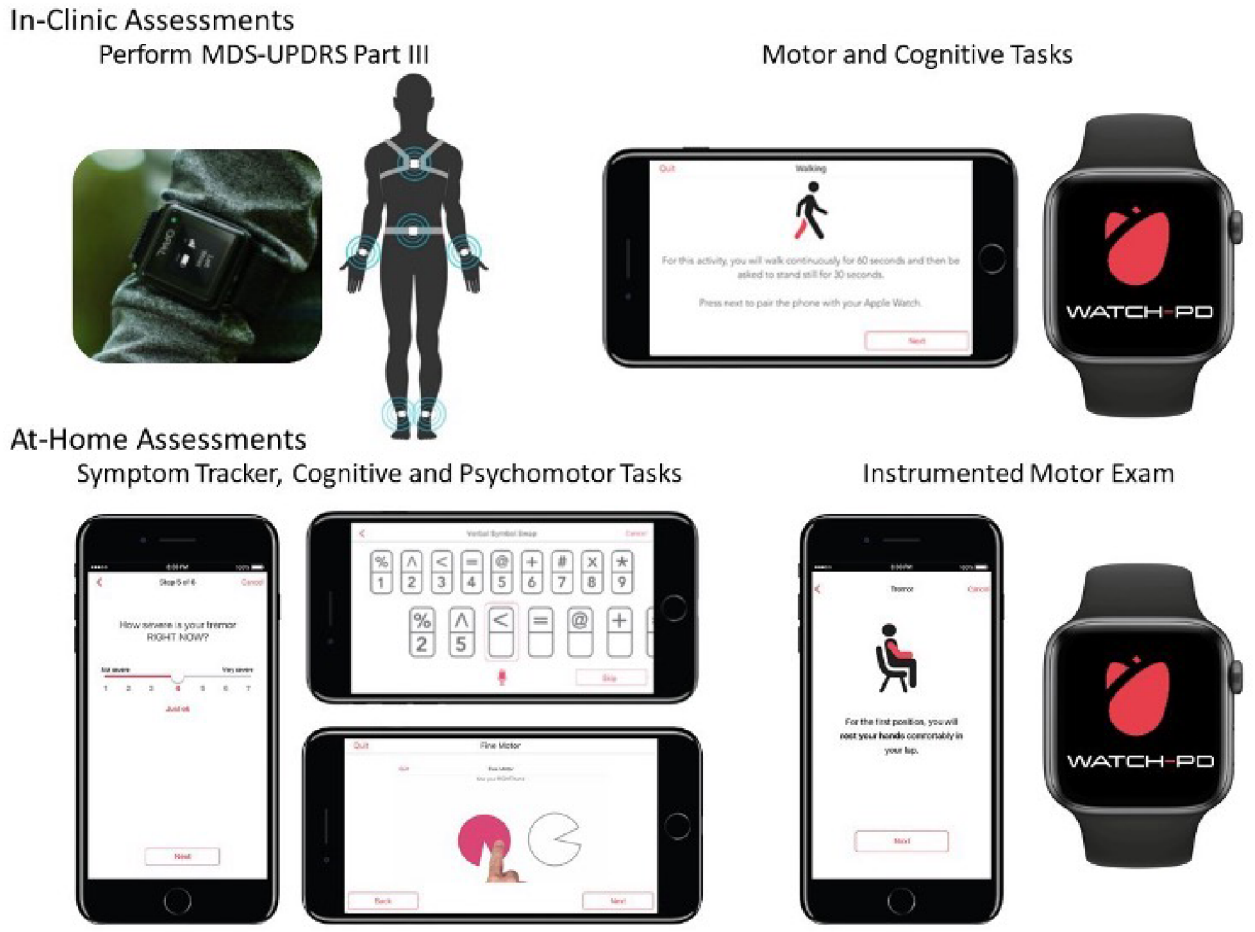


FIGURE 1
Digital devices evaluated in-clinic and at-home during WATCH-PD. Adapted from Adams et al. (4).

mobile battery, developed by Clinical Ink (Clinical Ink, Horsham, PA USA), at each clinic visit and every two weeks during their participation in the study. The purpose of completing the mobile battery during clinic visits was twofold. First, it provided a means of orienting participants to the devices and tasks to be performed. In addition, it allowed comparison of performance measures derived from the mobile assessments to contemporaneously collected measures acquired through the Opal system and clinician and patient reported outcomes completed during each visit. The mobile battery took approximately 15–20 min to complete, and participants were asked to complete the entire battery at once. They were allowed up to an hour to complete the tasks, providing time for unexpected interruptions or breaks. The battery consisted of three core components, measuring both motor and non-motor domains. First, participants completed a set of six, brief PRO questions, providing responses on a 1–7 Likert scale with questions related to current mood, fatigue, sleepiness, and cognition, as well as the current severity of bradykinesia and tremor (Table 1). Participants then completed a set of brief cognitive and psychomotor tasks and a brief speech recording battery. Finally, participants completed a brief instrumented motor exam consisting a of a 1-minute walking task, a 30 s balance task, and 20 s resting and postural tremor tasks.

2.6 Mobile device proficiency questionnaire

At baseline participants completed an abridged version of the MDPQ focused on a subscale of Mobile Device Basics most relevant to the tools being used in the current study. The MDPQ Mobile Device Basics subscale is comprised of nine questions that ask participants to rate their ability to perform tasks on a smartphone or tablet device on a 1–5 Likert scale (1 = never tried, 2 = not at all, 3 = not very easily, 4 = somewhat easily, 5 = very easily). The MDPQ was available for all participants at Baseline.

2.7 Wearability and comfort questionnaire

At baseline, Months 1, 6, and end of study (month 12), participants were asked to take a questionnaire with quantitative questions related to using the digital technologies both in the clinic and at home (Supplemental 1). Quantitative questions relating to comfort, ease of use, and burden were either on a Likert or categorical (Yes/No/Neutral) scale. The Likert Scale was a 1 to 5 scale for both comfort of devices (1 = Very Acceptable, 2 = Acceptable, 3 = Neutral, 4 = Unacceptable, 5 = Very Unacceptable) and ease of use (1 = Very easy, 2 = Easy, 3 = Neutral, 4 = Difficult, 5 = Very Difficult). At the end of the study, participants completed an exit questionnaire which addressed qualitative questions related to the use of the devices, and non-device questions related to length of study and compensation.

At baseline, the Wearability and Comfort Questionnaire was available for 80 participants with PD and 49 controls, however some questions were left blank which is reflected in our results. At month 1, the Wearability and Comfort Questionnaire was available for 72 participants with PD and 40 controls and at month 12, it available for 80 participants with PD and 46 controls.

TABLE 1 Results of the mobile device proficiency questionnaire (MDPQ) in Parkinson's disease participants and controls at baseline.

	PD (N = 82)	Control (N = 50)	P-value
MDPQ score			
Mean (SD)	20.6 (2.91)	21.5 (2.94)	0.0962
Median [Min, Max]	20.0 [15.0, 25.0]	20.5 [15.0, 25.0]	

2.8 Statistical analysis

Descriptive statistics for the MDPQ total subset score and Wearability and Comfort Questionnaire scores at Baseline, Month 1, and Month 12 were reported for PD participants and controls. A two-tailed t-test was performed between PD participants and controls on the MDPQ to determine if there was a difference in scores between the two groups where $p < 0.05$ was considered statistically significant. All analyses were performed in R Statistical Software (v4.1.2; R Core Team 2021).

3 Results

3.1 Mobile device proficiency questionnaire

Table 2 summarizes the results of the MDPQ at Baseline. The mean [SD] score in PD participants [20.6 (2.91)] was numerically smaller than controls [21.5 (2.94)] but did not differ significantly across the two groups ($p = 0.10$).

3.2 Wearability and comfort questionnaire

Figure 2 summarizes the results of the Wearability and Comfort Questionnaire at Baseline, Month 1, and Month 12.

3.3 Baseline

For overall comfort of the devices, the majority of the PD participants (75.9%) found the comfort of wearing the Opals to be very acceptable. Positive feedback was also reported for the mobile phone and smartwatch with the majority of participants reporting the comfort of the devices very acceptable (71.2% and 78.8% respectively). Similarly, controls reported very acceptable comfort for the Opals (77.6%), mobile phone (79.6%), and smartwatch (77.6%).

In relation to the mobile assessment, 85.2% of PD participants found the instructions on the mobile assessment to be clear and easy to understand, but 81.5% found the text was not easy to read. Likewise, 87.8% of controls found the instructions on the mobile assessment easy to be clear and easy to understand and 79.6% found the text was not easy to read (Figure 3).

TABLE 2 Description and location of assessments conducted with the digital devices used in WATCH-PD[a].

Device	Assessment	Description	Location
Wearable sensors	Timed walk test	The participant is timed while walking for a distance of 10 meters. The individual walks the 10-m path back and forth, turning at the end of their path, for 2 min.	In-clinic
	Repeat timed walk test with serial sevens	The participant repeats the Timed Walk Test described above. While walking, the participant performs a serial subtraction of sevens beginning with the number 100.	
	Sit-to-stand test	The participant sits against the back of a chair and stands up as quickly as they can for 5 repetitions without stopping.	
	Postural sway	The participant stands still, looking straight ahead for 30 s.	
Smartphone application	Symptom tracker	The participant answers a 5-item survey on the phone including questions about mood, sleepiness, thinking, tremor severity, and difficulty with movement.	In-clinic and remotely biweekly
	Symbol digit modalities test	The participant is given a key that connects symbols to numbers. The participant is presented with a symbol and must speak aloud the corresponding number.	
	Trail making test	The participant must connect a set of dots as quickly as possible using the index finger on their dominant hand while still maintaining accuracy.	
	Visuospatial working memory	The participant is briefly shown four colored boxes. The participant is then shown a single-colored box and must indicate if that box was in the previous set of four.	
	Finger tapping	The participant performs rapid alternating finger movements by tapping two targets that appear side by side using their index and middle fingers.	
	Fine motor test	The participant is presented with a pink object and an outline. The individual must use 1–2 fingers to move and rotate the object into the outline as quickly as possible.	
	Speech assessment	Participants must perform a sustained phonation task, a verbal articulation task repeating the syllables "pa ta ka," and a sentence reading task.	
Smartwatch	Timed walk test	The participant must walk in a straight line, turning at the end of their path, for 1 min.	In-clinic and remotely biweekly
	Balance test	The participant must stand still with their arms at their side for 30 s.	
	Tremor task	The participant must rest their hands in their lap for 10 s, then extend their arms out in front of them for 10 s.	

[a]Table 2 adapted from Adams et al. (4).

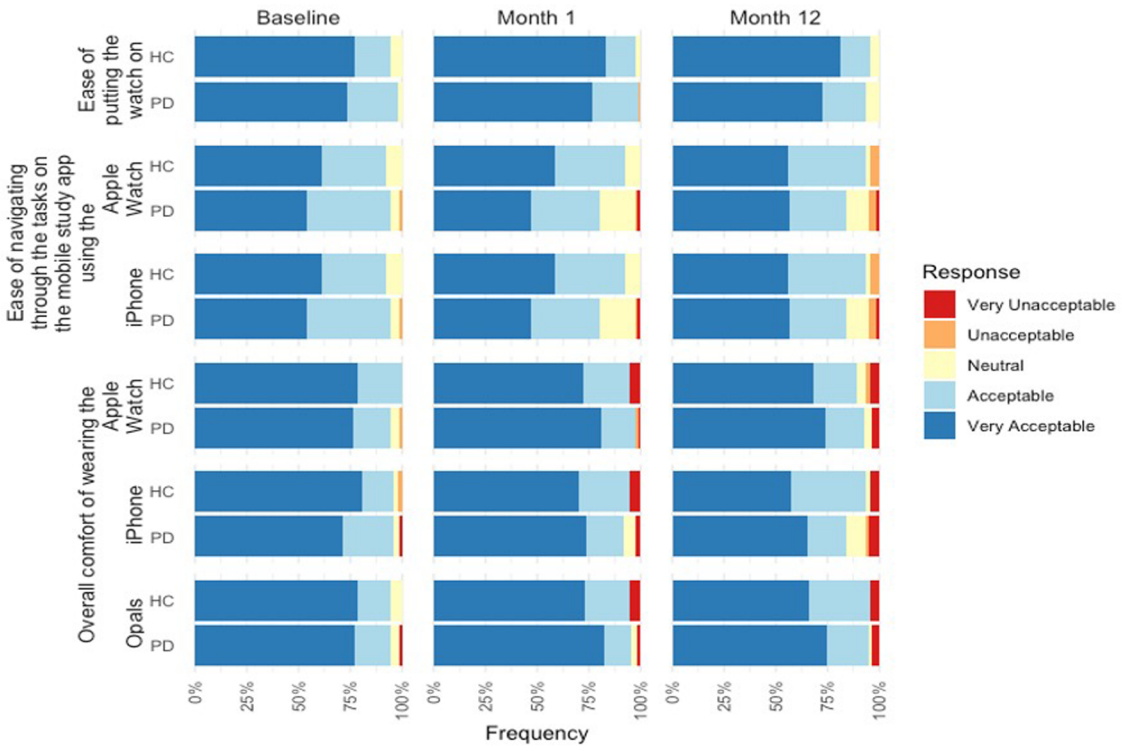


FIGURE 2
Impressions of using DHTs at baseline, Months 1, and 12 of PD and control participants.

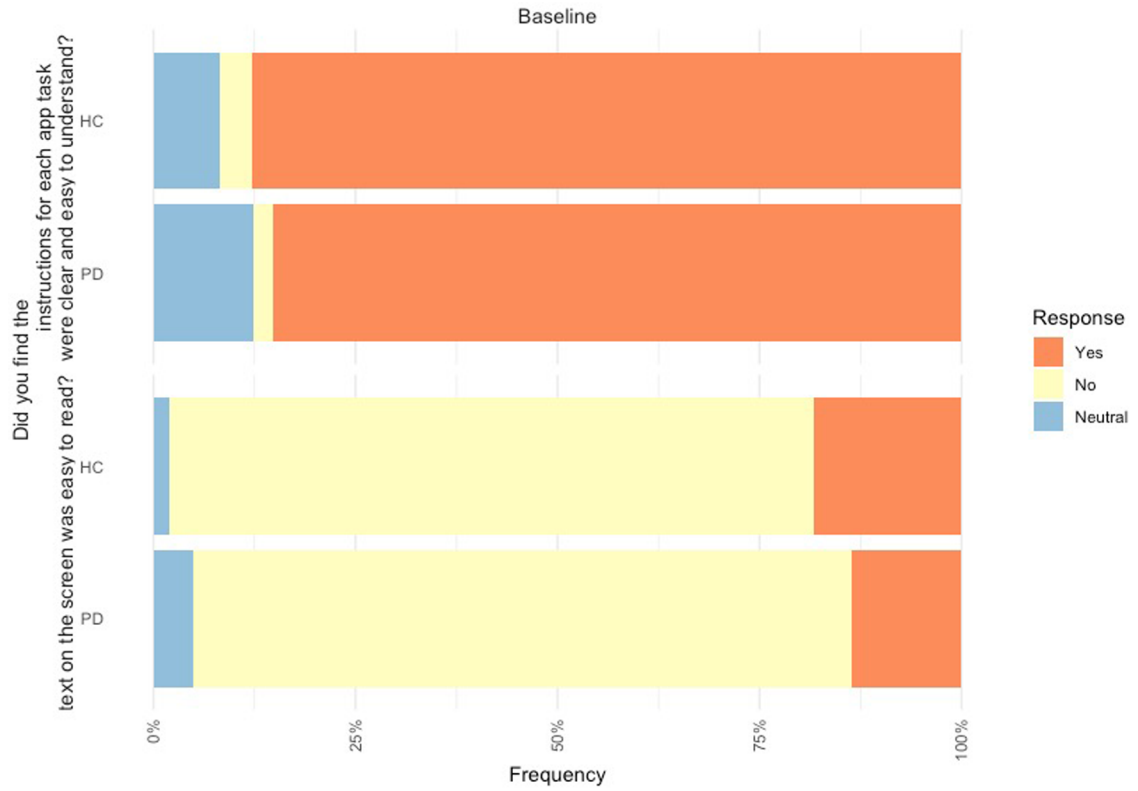


FIGURE 3
Feedback of mobile device platform at baseline from PD participants (n = 80) and controls (n = 49).

3.4 Month 1

For overall comfort of the devices in Month 1, the majority of the PD participants (82.1%) found the comfort of wearing the Opals to be very acceptable. Positive feedback was also reported for the mobile phone and smartwatch with the majority of participants reporting the comfort of the devices very acceptable (74.0% and 80.8% respectively), including acceptability of putting the smartwatch on at home (76.7%). Similarly, controls reported very acceptable comfort for the Opals (73.0%), mobile phone (70.0%), and smartwatch (72.5%), and 83.0% reported the ease of putting on the smartwatch as very acceptable.

3.5 Month 12

For overall comfort of the devices in Month 12, the majority of PD participants (74.4%) found the comfort of wearing the Opals to be very acceptable. Positive feedback was also reported for the mobile phone and smartwatch with the majority of participants reporting the comfort of the devices very acceptable (65.0% and 73.8% respectively), including acceptability of putting the smartwatch on at home (72.5%). Similarly, controls reported very acceptable comfort for the Opals (66.0%), mobile phone (57.4%), and smartwatch (68.1%), and 83.0% reported the ease of putting on the smartwatch as very acceptable.

Highlights of the qualitative feedback related to the devices at Month 12 was grouped and can be found in Table 3. Participants highlighted the need for a better watch strap, more notifications on the mobile device to complete the battery, and frustrations with technological issues.

TABLE 3 Qualitative feedback from participants on use of smartwatch and smartphone at-home in WATCH-PD.

Smartwatch	Smartphone
"Too bulky" *"Sometimes when tremors are acting up, the watch was uncomfortable"* *"The strap interfered with writing and using a computer mouse"* *"Would prefer to wear it on non-dominant wrist"* *"The walking activity, when both the phone and watch are used alternatively, was confusing"* *"Would be helpful to have study reminders on the watch"*	*"The study phone was just a brick when not used for sessions"* *"Sometimes between uses the battery dies so a call or text would be helpful"* *"A call or text on my personal phone to remind me about study tasks would be helpful"* *"Froze once so I had to reboot"*

4 Discussion

This work aimed to gather participant perceptions of the DHTs used in the WATCH-PD study. This is the first study to evaluate feedback and impressions of using common DHTs in both controls and people with early PD in the context of a longitudinal, observational study. We show that for an early PD

population, experiences and comfort with technology are not different from the general experience in neurologically healthy older adults. Furthermore, there was an overall favorable view of the usability and comfort of the digital technologies deployed in the WATCH-PD study, both in-clinic and at-home.

Results from the MDPQ mobile subscale at baseline demonstrated no significant differences in device proficiency between the PD participants and controls. The results from the Wearability and Comfort Questionnaire overall demonstrated generally positive views on the comfort and use of the digital technologies in this study. Consistently, over the 12-month study duration, within both cohorts, most participants found wearing the Opal sensors, mobile phone, and smartwatch either very acceptable or acceptable regarding comfort. The ease of putting on the Apple Watch band was also favorable throughout the study, which was encouraging given that many of the PD participants presented with tremor dominant symptoms at baseline.

The study is not without limitations. The baseline MDPQ scores combined with the highly positive results on the Wearability and Comfort Questionnaire might suggest that the study was biased towards recruiting people who were already very comfortable with technology. This cohort was also homogenous, potentially limiting the generalizability of our findings. Thus, it is recommended that future work collect similar measures in more diverse cohorts, potentially through a fully remote study design to widen recruitment and include a broader range of individuals. Moreover, there were a few limitations which we could not control, including the maximum size of the screen of the mobile device.

5 Conclusions

The current research in early PD, along with extant literature on DHT usability and acceptability more generally, provides a foundation for understanding the acceptability of using digital tools in early PD clinical trials. Our work provides insights into how older individuals, especially those with a movement disorder, will adapt to using digital technologies in clinical trials. A key to overcoming possible challenges with the use of DHTs in older participants with neurological disorders is to incorporate the patient voice by gathering regular formal and informal feedback throughout study design and conduct. Furthermore, the option of co-design with the end users provides an opportunity to collect valuable feedback and create a collaborative experience between researchers and patients.

Data availability statement

The datasets presented in this article are not readily available because data is available to members of the Critical Path for Parkinson's Consortium 3DT Initiative Stage 2. For those who are not a part of 3DT Stage 2, a proposal may be made to the WATCH-PD Steering Committee (via the corresponding author) for de-identified datasets. Requests to access the datasets should be directed to Jamie.Adams@chet.rochester.edu.

Ethics statement

The studies involving humans were approved by WCG Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

TK: Writing – original draft, Writing – review & editing. RL: Writing – review & editing. JA: Writing – review & editing. RD: Writing – review & editing. MK: Writing – review & editing. JS: Writing – review & editing. DA: Writing – review & editing. FH: Writing – review & editing. DS: Writing – review & editing. JC: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Funding was provided by Takeda, Biogen, and Critical path Institute.

Acknowledgments

We want to thank the members of the Critical Path for Parkinson's Digital Drug Development tool (3DT) initiative for generously funding and collaborating to support the 3DT project and a talented group of scientific advisors who have supported this effort. Members include Biogen, Takeda, AbbVie, Novartis, Merck, UCB, GSK, Roche, The Michael J. Fox Foundation and Parkinson's UK.

Conflict of interest

The 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

References

1. Sidoti O. *Mobile Fact Sheet*. Washington, DC: Pew Research Center (2024). Available online at: https://www.pewresearch.org/internet/fact-sheet/mobile/

2. Vogels E. *About One-in-Five Americans Use a Smart Watch or Fitness Tracker*. Washington, DC: Pew Research Center (2020). Available online at: https://www.pewresearch.org/short-reads/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/

3. Steinhubl SR, Wolff-Hughes DL, Nilsen W, Iturriaga E, Califf RM. Digital clinical trials: creating a vision for the future. *NPJ Digital Medicine*. (2019) 2:126. doi: 10.1038/s41746-019-0203-0

4. Adams JL, Kangarloo T, Tracey B, O'Donnell P, Volfson D, Latzman RD, et al. Using a smartwatch and smartphone to assess early Parkinson's disease in the WATCH-PD study. *NPJ Parkinson's Dis*. (2023) 9(1):64. doi: 10.1038/s41531-023-00497-x

5. Stroud C, Onnela J-P, Manji H. Harnessing digital technology to predict, diagnose, monitor, and develop treatments for brain disorders. *NPJ Digit Med*. (2019) 2:44. doi: 10.1038/s41746-019-0123-z

6. 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

7. Stephenson D, Badawy R, Mathur S, Tome M, Rochester L. Digital progression biomarkers as novel endpoints in clinical trials: a multistakeholder perspective. *J Parkinson's Dis*. (2021) 11:S103. doi: 10.3233/jpd-202428

8. Czech MD, Badley D, Yang L, Shen J, Crouthamel M, Kangarloo T, et al. Improved measurement of disease progression in people living with early Parkinson's disease using digital health technologies. *Commun Med*. (2024) 4(1):49. doi: 10.1038/s43856-024-00481-3

9. Dodge HH, Zhu J, Mattek NC, Austin D, Kornfeld J, Kaye JA. Use of high-frequency in-home monitoring data may reduce sample sizes needed in clinical trials. *PloS One*. (2015) 10:e0138095. doi: 10.1371/journal.pone.0138095

10. Brouwer WD, Patel CJ, Manrai AK, Rodriguez-Chavez IR, Shah NR. Empowering clinical research in a decentralized world. *npj Digit Med*. (2021) 4 (1):102. doi: 10.1038/s41746-021-00473-w

11. Chen C, Ding S, Wang J. Digital health for aging populations. *Nat Med*. (2023) 29(7):1623–30. doi: 10.1038/s41591-023-02391-8

12. Gitlow L. Technology use by older adults and barriers to using technology. *Phys Occup Ther Geriatr*. (2014) 32(3):271–80. doi: 10.3109/02703181.2014.946640

13. Moxley J, Sharit J, Czaja SJ. The factors influencing older Adults' decisions surrounding adoption of technology: quantitative experimental study. *JMIR Aging*. (2022) 5:e39890. doi: 10.2196/39890

14. Samii A, Nutt JG, Ransom BR. Parkinson's disease. *Lancet (London, England)*. (2004) 363(9423):1783–93. doi: 10.1016/S0140-6736(04)16305-8

15. Balestrino R, Schapira AHV. *Parkinson disease*. *Eur J Neurol*. (2020) 27 (1):27–42. doi: 10.1111/ene.14108

16. Port RJ, Rumsby M, Brown G, Harrison IF, Amjad A, Bale CJ. People with Parkinson's disease: what symptoms do they most want to improve and how does this change with disease duration? *J Parkinsons Dis*. (2021) 11(2):715–24. doi: 10.3233/JPD-202346

17. 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) 23(15):2129–70. doi: 10.1002/mds.22340

18. Dorsey ER, Papapetropoulos S, Xiong M, Kieburtz K. The first frontier: digital biomarkers for neurodegenerative disorders. *Digit Biomark*. (2017) 1(1):6–13. doi: 10.1159/000477383

19. Simuni T, Siderowf A, Lasch S, Coffey CS, Caspell-Garcia C, Jennings D, et al. Longitudinal change of clinical and biological measures in early Parkinson's disease: parkinson's progression markers initiative cohort. *Mov Disord*. (2018) 33:771–82. doi: 10.1002/mds.27361

20. Vaswani PA, Tropea TF, Dahodwala N. Overcoming barriers to Parkinson disease trial participation: increasing diversity and novel designs for recruitment and retention. *Neurotherapeutics*. (2020) 17:1724–35. doi: 10.1007/s13311-020-00960-0

21. Roque NA, Boot WR. A new tool for assessing Mobile device proficiency in older adults: the Mobile device proficiency questionnaire. *J Appl Gerontol*. (2018) 37:131–56. doi: 10.1177/0733464816642582

22. Marek K, Jennings D, Lasch S, Siderowf A, Tanner C, Simuni T, et al. The Parkinson progression marker initiative (PPMI). *Biol Mark Neurodegenerative Dis*. (2011) 95(4):629–35. doi: 10.1016/j.pneurobio.2011.09.005

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 09 October 2024
DOI 10.3389/fdgth.2024.1430994

OPEN ACCESS

EDITED BY
Max A. Little,
University of Birmingham, United Kingdom

REVIEWED BY
Siddharth Arora,
University of Oxford, United Kingdom
Marjan J. Meinders,
Radboud University Medical Centre,
Netherlands

*CORRESPONDENCE
Jennie S. Lavine
✉ jennie.lavine@koneksahealth.com
Larsson Omberg
✉ larsson.omberg@koneksahealth.com

RECEIVED 11 May 2024
ACCEPTED 12 September 2024
PUBLISHED 09 October 2024

CITATION
Lavine JS, Scotina AD, Haney S, Bakker JP, Izmailova ES and Omberg L (2024) Impacts on study design when implementing digital measures in Parkinson's disease-modifying therapy trials.
Front. Digit. Health 6:1430994.
doi: 10.3389/fdgth.2024.1430994

Impacts on study design when implementing digital measures in Parkinson's disease-modifying therapy trials

Jennie S. Lavine*, Anthony D. Scotina, Seth Haney, Jessie P. Bakker, Elena S. Izmailova and Larsson Omberg*

Research & Development, Koneksa Health, New York, NY, United States

Introduction: Parkinson's Disease affects over 8.5 million people and there are currently no medications approved to treat underlying disease. Clinical trials for disease modifying therapies (DMT) are hampered by a lack of sufficiently sensitive measures to detect treatment effect. Reliable digital assessments of motor function allow for frequent at-home measurements that may be able to sensitively detect disease progression.
Methods: Here, we estimate the test-retest reliability of a suite of at-home motor measures derived from raw triaxial accelerometry data collected from 44 participants (21 with confirmed PD) and use the estimates to simulate digital measures in DMT trials. We consider three schedules of assessments and fit linear mixed models to the simulated data to determine whether a treatment effect can be detected.
Results: We find at-home measures vary in reliability; many have ICCs as high as or higher than MDS-UPDRS part III total score. Compared with quarterly in-clinic assessments, frequent at-home measures reduce the sample size needed to detect a 30% reduction in disease progression from over 300 per study arm to 150 or less than 100 for bursts and evenly spaced at-home assessments, respectively. The results regarding superiority of at-home assessments for detecting change over time are robust to relaxing assumptions regarding the responsiveness to disease progression and variability in progression rates.
Discussion: Overall, at-home measures have a favorable reliability profile for sensitive detection of treatment effects in DMT trials. Future work is needed to better understand the causes of variability in PD progression and identify the most appropriate statistical methods for effect detection.

KEYWORDS

Parkinson's disease, digital health technology, measurement reliability, clinical trials, statistical power, disease progression, longitudinal data, simulation study

1 Introduction

Parkinson's Disease (PD) is a slow-progressing neurodegenerative disease that affects over 8.5 million people worldwide and is currently the fastest growing neurodegenerative disease in the world (1). Hallmarks of PD include slowness of movement and rigidity, and the impacts are felt in many aspects of everyday motor function including gait, eating, speech, and dressing. Currently available PD medications address symptoms but do not treat the underlying disease. Recent advances in drug development show promise for disease modifying therapies (DMTs) but evaluation of these treatments is hampered by outcome measures such as the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS), which requires large sample sizes and/or long

term follow-up to detect modest treatment effects, especially given that existing symptomatic treatment can mask underlying progression (2). Digital at-home measures, which allow for more frequent assessment, are a promising option for detecting treatment effects in shorter timeframes and/or with a smaller number of participants.

Digital measures are currently recommended as exploratory endpoints in randomized controlled trials (RCTs) (3). For use as primary and secondary endpoints, and regardless of whether the measure is considered a biomarker or a clinical outcome assessment, a better understanding of their reliability and responsiveness to disease progression is necessary to determine their optimal context of use and assessment schedule. Clinimetric properties of digital tools have been assessed in a wide range of studies to determine how they can be useful in PD (see Supplementary Table 1 and references within). Multiple studies of digital measures derived from at-home app-based assessments, such as finger tapping and timed walk tests, demonstrate associations with aligned in-clinic assessments and high test-retest reliability [(4–7), Supplementary Table 1]. The reliability of many of these measures is as good as or better than test-retest reliability for MDS-UPDRS part III scores (8).

In current clinical trials for novel DMTs for PD, the MDS-UPDRS or one of its subparts is the gold standard outcome measure (3). Composed of four parts, each of which consists of multiple items scored ordinally from 0 to 4 (where 0 is no symptoms and 4 is severe symptoms), the items comprise patient-reported outcomes and clinician assessments (9). Parts II and III relate to motor function, measuring patient perception and clinician ratings of motor impacts respectively. These parts have excellent test-retest reliability as measured by intraclass correlation coefficients (ICCs) across spans of 1−2 weeks [ICCs for part II: 0.96, part III: 0.93 (8)]; however, it remains challenging to detect changes in early disease burden, especially in the face of symptomatic treatments (2). One explanation for this apparent conundrum is that there are three fundamentally different sources of variability in measurements of PD motor function: measurement error, short-term clinical fluctuations, and long-term variability in underlying disease progression.

On the timescale of days to a few weeks, there is no expectation of change in underlying disease severity, yet measures vary from one time point to the next due to measurement error and day-to-day fluctuations in symptoms. Measurement error may be present in clinician ratings due to, for example, interrater reliability (10, 11) and in at-home digital assessments due to, for example, variability in the setting in which patients use the digital devices assessments (12). Also on a short time scale, clinical variability results from day-to-day and diurnal symptom fluctuations including those induced by levodopa and other symptomatic treatment medications (13). These types of variability can be quantified with the ICC, standard error of measurement (σ_m), and minimum detectable change (MDC) in cross-sectional studies and have been established for both in-clinic and at-home assessments.

In contrast, long-term variability in underlying disease progression arises from PD being a heterogeneous disease. When averaged over individuals, the progression of PD motor manifestations as measured by MDS-UPDRS or digital assessments can be approximated as linear over the span of a year or two (2, 14). However, PD's motor manifestations do not change at a constant rate across months within (2, 15) or between (16, 17) individuals. The causes of inter- and intra-individual variability in disease progression are not well known and may include differences in underlying disease etiology, seasonality, stress, climate, and changes in living situation (15, 16). Variability in progression rate is harder to estimate because it is only apparent at long timescales; however, it is detectable in longitudinal MDS-UPDRS data such as those collected in the PPMI study (18) and has been disentangled from measurement error by Evers et al. (15).

Digital assessments can help overcome the challenges posed to clinical trials by all three of these types of variability by allowing for more frequent measures. Including repeat measures reduces the standard error of endpoint estimates such as the rate of change from baseline. In contrast with clinician-observed outcome assessments, which are typically captured infrequently due to the burden and cost of clinic visits, the schedule of assessments for digital measures can be driven by study designs that yield the highest power for detecting the treatment effect.

Multiple outcome measures have been considered from assessments completed using digital tools. These include individual measures, such as number of taps or gait speed derived from a mobile app-based assessment, and summary statistics of a burst of the same assessment, such as the median of 6 tapping assessments completed over the course of seven days. There is a trade-off between these two outcome measures: individual measures can be completed more frequently, but median values of bursts have higher test-retest reliability (7).

While digital measures have been used in clinical trials as exploratory endpoints, it remains unclear under what conditions they will outperform in-clinic assessments and how best to distribute assessments across the length of the trial to detect the treatment effect. We undertook analyses to address these gaps with the following objectives: (1) Estimate measurement error in a variety of at-home digital assessments spanning gait, tapping, and tremor, which are part of a neuroscience toolkit developed by Koneksa Health for use in clinical trials. The measures, derived from raw triaxial accelerometry sensor data (19), were applied to data collected in the Objective PD sub-study of the mPower study (20). (2) Simulate various DMT study designs that implement individual measures and bursts using at-home digital assessments vs. in-clinic MDS-UPDRS. We use the Gaussian state space framework developed by Evers et al. (15), which explicitly models measurement error and variability in disease progression rates. (3) Assess the power to detect a treatment effect in the various scenarios by fitting linear mixed effects models to the simulated measures.

2 Methods

2.1 Data

The data used in this study to estimate reliability of digital assessments derive from the ObjectivePD sub-study (20), which

recruited 44 participants (21 with confirmed PD diagnosis, 23 healthy controls). Participants were followed for 6 months and seen in clinic three times at 0, 3 and 6 months. During the entire 6 months, they were also asked to complete daily digital health measures administered through the mPower mobile application (20). These assessments consisted of (1) speeded finger tapping alternating between the index and middle finger, (2) a 30-s walk test with the phone in the pocket, and (3) three tremor assessments including resting, postural and hand-to-nose tremor. Each participant in the ObjectivePD sub-study performed on average 182 tapping sessions, 147 gait assessments, and 134 tremor sessions throughout the 6 months study timeframe. Additional details of the measures are available in prior publications (20, 21) and Supplementary Table 2.

2.2 Reliability measure estimation

We estimated measurement error and test-retest reliability of at-home digital measures using a linear random intercept model. We assessed the test-retest reliability of measures derived from individual at-home assessments and measures that summarize multiple at-home assessments completed within a 7-day period with their median. Specifically, at-home measurements assessed longitudinally per participant were grouped by fortnight, and a linear model was fit per digital measure with random intercepts for participant and participant-by-fortnight interaction. In contrast with conventional methods for calculating test-retest reliability that rely on two parallel assessments (e.g., assessments taken on the same participant over a short period of time, or assessments collected from two raters at the same point in time), assessment of test-retest reliability with a longitudinal model uses all measurements collected during the study and are robust to missing data (22). Furthermore, test-retest reliability in this context can be interpreted as the consistency between measurements collected during any 2-week period. Implicit in this calculation is the assumption that underlying disease progression between observations within a fortnight will be minor (8). This analysis was performed separately for measures that summarized bursts and measures that represented individual assessments. Model residuals were plotted to assess whether the model was an appropriate choice.

For each fitted model, we extracted the measurement error associated with a particular measure as the residual variance, σ^2_m. Test-retest reliability, assessed with the intraclass correlation coefficient, is extracted from the fitted model; it is the proportion of the overall variability in a digital measure explained by the participant effect and the participant-by-fortnight interaction effect.

We calculated the minimum detectable change (MDC) associated with each digital measure following Weir (23) as:

$$\mathrm{MDC} = 1.96 \times \sqrt{2} \times \sigma_m$$

2.3 Model for simulating digital and in-clinic data

We generated simulated study data from a Gaussian state space model of PD progression and measurement (Figure 1) that showed a good fit to longitudinal MDS-UPDRS data from the PPMI cohort (15), see Supplementary Text for further discussion of the modeling framework). In brief, unobserved underlying disease severity, θ, is simulated for a study population of size n by randomly drawing n initial values from a normal distribution. Each participant's disease severity is updated to the next time step by adding the mean trend, τ, (i.e., the underlying disease progression rate) plus Gaussian noise representing variability in the progression process (σ_T). The rate of disease progression, τ, is the only parameter that differs between placebo and DMT study arms. The updating procedure is repeated for each participant across the length of a simulated trial with Q observed timepoints. Observed values, y, are then simulated from the time series of underlying disease severity, θ, by adding normally distributed measurement error, v, representing a combination of inter- and intra-rater reliability and short-term fluctuations that are not related to underlying disease progression. The updating process is encapsulated in the following equations, for $i \in \{1, 2, \ldots n\}$ and $t \in \{1, 2, \ldots Q\}$.

$$y_{t,i} = \theta_{t,i} + v_{t,i},\ v_{t,i} \sim N(0,\ \sigma_m)$$

$$\theta_{t,i} = \theta_{t-1,i} + w_{t,i},\ w_{t,i} \sim N(\tau,\ \sigma_T)$$

The elements of clinical study design included in the simulations were the number of participants per study arm, schedule of assessments, and study duration. For simulations of in-clinic MDS-UPDRS part III scores, all parameters were taken from estimates described in Evers et al. (15).

For simulations of digital at-home measures, measurement parameters were estimated from the mPower data (i.e., starting mean, μ_s, starting standard deviation, σ_s and standard error of the measurement, σ_m, as described above for individual assessments). Bursts were simulated by drawing 6 individual assessments per burst and taking the median. Unfortunately, we lack empirical estimates of the trend and trend variance (τ and σ^2_T) from at-home assessments because we do not have sufficient longitudinal data on digital measures to disentangle measurement error from progression variability.

Because τ and σ^2_T represent the trend and trend variance in underlying disease progression, respectively, we began by assuming that these are independent of measurement type and scale with the mean value of a measure, which allowed us to estimate them from the in-clinic measures. That is, $\tau_{\mathrm{digital}} = \tau_{\mathrm{clinic}}$ $(\mu_{\mathrm{digital}}/\mu_{\mathrm{clinic}})$ and $\sigma_{\mathrm{T,digital}} = \sigma_{\mathrm{T,clinic}}$ $(\mu_{\mathrm{digital}}/\mu_{\mathrm{clinic}})$. However, while in-clinic and at-home assessments both measure underlying motor function, they do so in somewhat different ways, and we therefore relaxed this assumption and considered the robustness of our results to the possibility that at-home measures may be less responsive than in-clinic measures by

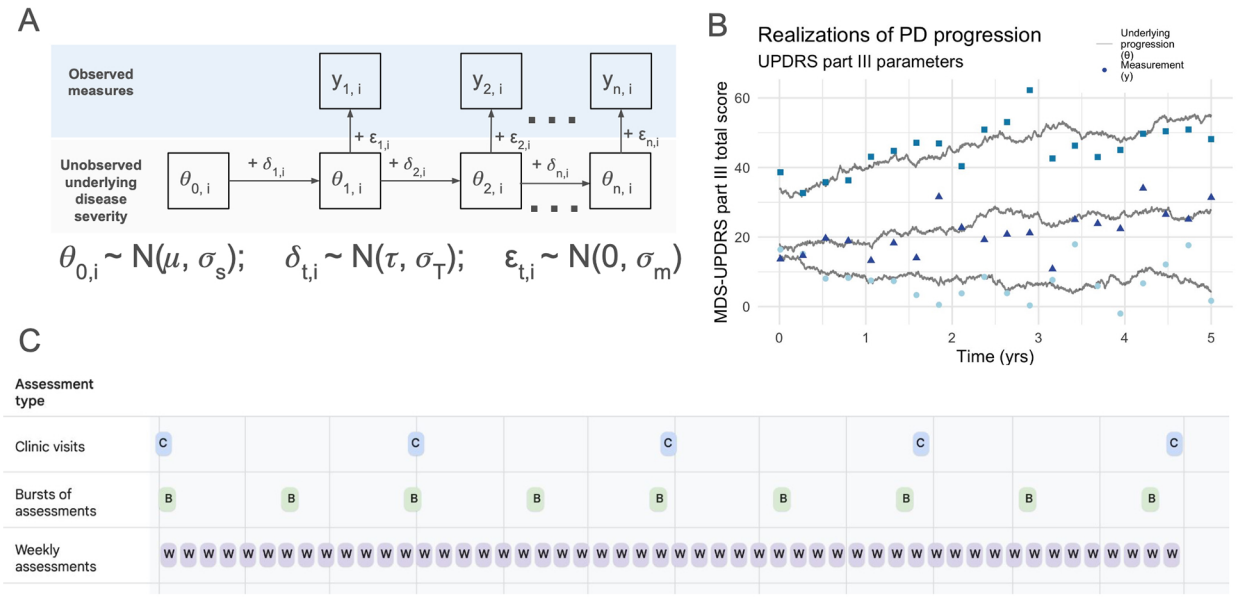


FIGURE 1
Conceptual model and simulation framework. **(A)** Visual representation of the Gaussian state space model used for simulations. **(B)** Three stochastic realizations of the model, using in-clinic MDS-UPDRS part III parameters (see Table 1). Underlying disease progression is represented by gray lines in **(B)** and $\theta_{t,i}$ in **(A)** The variability in progression rates between individuals and across time within individuals arises from the variability in the trend, simulated by $\delta_{t,i}$ and result in the unobserved underlying disease states (gray lines). The observed measurements (i.e., MDS-UPDRS scores, digital assessment scores, etc.) are represented by $y_{t,i}$ in **(A)** and points in **(B)** The vertical distance between the gray line and its associated points represents the variability induced by the measurement process, $\varepsilon_{t,i}$. **(C)** The three main study designs considered: (1) quarterly clinic visits, (2) bursts of 6 assessments 8 times per year, and (3) weekly individual assessments.

reducing $\tau_{digital}$ to varying degrees relative to in-clinic measures. We modeled the effect, e, of a DMT as a reduction in the progression rate, τ, such that the progression rate in the treatment arm is $e\tau$, where $0 < e \leq 1$.

2.4 Study designs

We considered three core study designs (Figure 1C): (1) in-clinic MDS-UPDRS every 3 months, (2) 48 digital at-home assessments per year clustered into 8 bursts of six assessments each, and (3) 48 digital at-home assessments per year evenly spaced across the study duration. We additionally assessed the robustness of our results to study designs with different clustering of bursts by grouping the 48 assessments into 4, 6, 12 and 24 bursts.

2.5 Progression rate estimation from simulations & power calculations

We used these simulations to determine the statistical power of a clinical trial to detect treatment effect. Statistical power measures the sensitivity of a study to an effect of interest and is used here to compare the sensitivity of different longitudinal study designs to detect reduced PD progression induced by a DMT. After simulating data for placebo and treatment arms, we assessed the power to detect a treatment effect by fitting a linear mixed effects model to the simulated observations, y, with fixed effects for time, study arm and their interaction, and a random intercept for participant. A first-order autoregressive, AR(1), process was used to model the residual covariance structure between observations within participants following model selection. Model residuals were examined to assess goodness of fit.

For every set of parameters, 1,000 simulations were run and statistical power was calculated as the proportion of assessments for which the coefficient of the interaction term for treatment-arm-by-time had a p-value <0.05, as determined from a t-distribution with the appropriate degrees of freedom using the R package *nlme* (24). An additional criterion for trial success is included in the supplement; in these simulations, in addition to a p-value <0.05, the mean difference in change between the treatment and placebo group across the study duration must exceed the MDC for the measure of interest.

2.6 Software

ICC calculations were performed using Python 3.11 and simulations and power calculations were carried out in R 4.2.1 (25). The code used for analysis and simulations is available upon request.

10.3389/fdgth.2024.1430994

TABLE 1 Parameters used in main simulations (Figure 2).

Parameter	MDS-UPDRS part III score [values from Evers (15)]	Digital at-home step length (meters)
τ	2.63 year^{-1} (13% year^{-1})	0.04–0.07 (8%–13% year^{-1})
σ_T	5.58 year^{-1}	0.15 year^{-1}
σ_m	3.94[a]	0.06[b]

[a]Independent estimate from Martinez-Martin (8): 4.3.
[b]Estimated from mPower data.

3 Results

3.1 Reliability of at-home digital PD assessments

We assessed the reliability of at-home digital measures obtained from (1) a 30-s walk test ("gait measures"), (2) a speeded finger tapping assessment ("finger tap measures"), and (3) a tremor assessment ("tremor measures"). Figure 2 summarizes the test-retest reliability as measured by ICCs for each at-home digital measure, separated by whether they were considered individually or an average across multiple measures taken within a 7-day period. Measurements obtained from bursts are summarized by calculating the median value per burst. A median of 6 measurements (mean = 4.6, standard deviation = 2.6) were included in each burst calculation. Most measures obtained from individual or burst assessments exhibited good-to-excellent reliability (26). However, several measures showed poorer reliability overall (e.g., log step time discrepancy, log tap interval symmetry, and tap correctness, collected during individual assessments; log tap interval change collected during burst assessments). The modeling approach used for estimation appeared reasonable based on Q-Q plots and other visualizations of residuals (Supplementary Figure 1). The MDC varied across measures, ranging from less than 10% of the mean (e.g., postural tremor displacement) to over 150% of the mean (e.g., change in tap interval) (Supplementary Tables 2, 3).

	MEDIAN BURST TEST-RETEST ICC	INDIVIDUAL ASSESSMENT TEST-RETEST ICC
GAIT MEASURES		
Step Period, sec	0.946	0.841
Stride Period, sec	0.946	0.841
Step Length, meters	0.938	0.803
Distance, meters	0.911	0.861
Gait Speed, meters/sec	0.910	0.804
Cadence, steps/min	0.905	0.797
Step Count	0.903	0.837
Z-axis Variability, g	0.873	0.538
Y-axis Variability, g	0.809	0.500
X-axis Variability, g	0.802	0.478
Freeze Index	0.801	0.587
Stride Similarity	0.797	0.698
XY-axis Variability, g	0.780	0.517
Log10(Step Time Discrepancy)	0.656	0.381
HAND-TO-NOSE TREMOR MEASURES		
Log10(Tremor Amplitude, g)	0.909	0.852
Peak Frequency, Hz	0.804	0.540
Adj. Peak Frequency, Hz	0.707	0.589
POSTURAL TREMOR MEASURES		
Log10(RMS Tremor Acceleration, g)	0.896	0.781
Tremor Frequency, Hz	0.892	0.677
Log10(Tremor Amplitude)	0.852	0.636
Log10(RMS Tremor Displacement, meters)	0.783	0.609
Peak Frequency Acceleration, Hz	0.751	0.676
RESTING TREMOR MEASURES		
Log10(RMS Tremor Acceleration, g)	0.800	0.623
Log10(RMS Tremor Displacement, meters)	0.768	0.549
Tremor Frequency, Hz	0.757	0.594
Log10(Tremor Amplitude)	0.699	0.548
Peak Frequency Acceleration, Hz	0.649	0.454
FINGER TAP MEASURES		
Tap Count	0.962	0.899
Tap Speed, taps/sec	0.944	0.865
Tap Positional Accuracy, pixels	0.741	0.641
Log10(Tap Regularity, sec)	0.652	0.419
Tap Correctness	0.632	0.351
Log10(Tap Interval Symmetry, sec)	0.512	0.285
Log10(Tap Interval Change, sec)	0.284	0.420

MDS-UPDRS Part II Test-Retest ICC: 0.96
MDS-UPDRS Part III Test-Retest ICC: 0.93
Reference: *Martinez-Martin, 2013*

FIGURE 2
Test-retest reliability per digital at-home measure and study design. For burst assessments, test-retest reliability is calculated between the median of measurements within each burst; for individual assessments, test-retest reliability is calculated between the individual measurements.

3.2 Power calculations for at-home measures & study design implications

Power calculations were carried out by fitting a linear mixed model to data generated from the Gaussian state space model. Examination of model residuals suggested a reasonable fit between the model used for effect detection and that used for data generation (Supplementary Figure 2). A comparison between mixed models with and without an autoregressive correlation structure of order 1 AR(1) indicated a significantly better fit by AIC values for the AR(1) model (Supplementary Figure 3), and that model is used for all power calculations presented here.

Based on empirical estimates of measurement error in digital and in-clinic assessments, and assuming that digital measures progress at the same rate as in-clinic measures after rescaling to account for different units, repeated at-home assessments consistently outperformed in-clinic assessments taken once every 3 months, regardless of whether the digital assessments were implemented in bursts or assessed weekly (evenly spaced), during a 1-year trial (Figure 3). For 2-year trials, at-home assessments implemented in bursts perform similarly to in-clinic assessments taken once every 3 months, assuming equivalent responsiveness.

As the responsiveness of digital measures (i.e., the trend, τ) decreases compared with clinic MDS-UPDRS Part III total score, statistical power decreases, regardless of the method of at-home assessment (collected weekly or within bursts). However, for the full range of parameters considered in these simulations, weekly at-home assessments retained higher statistical power compared to in-clinic assessments performed once every 3 months.

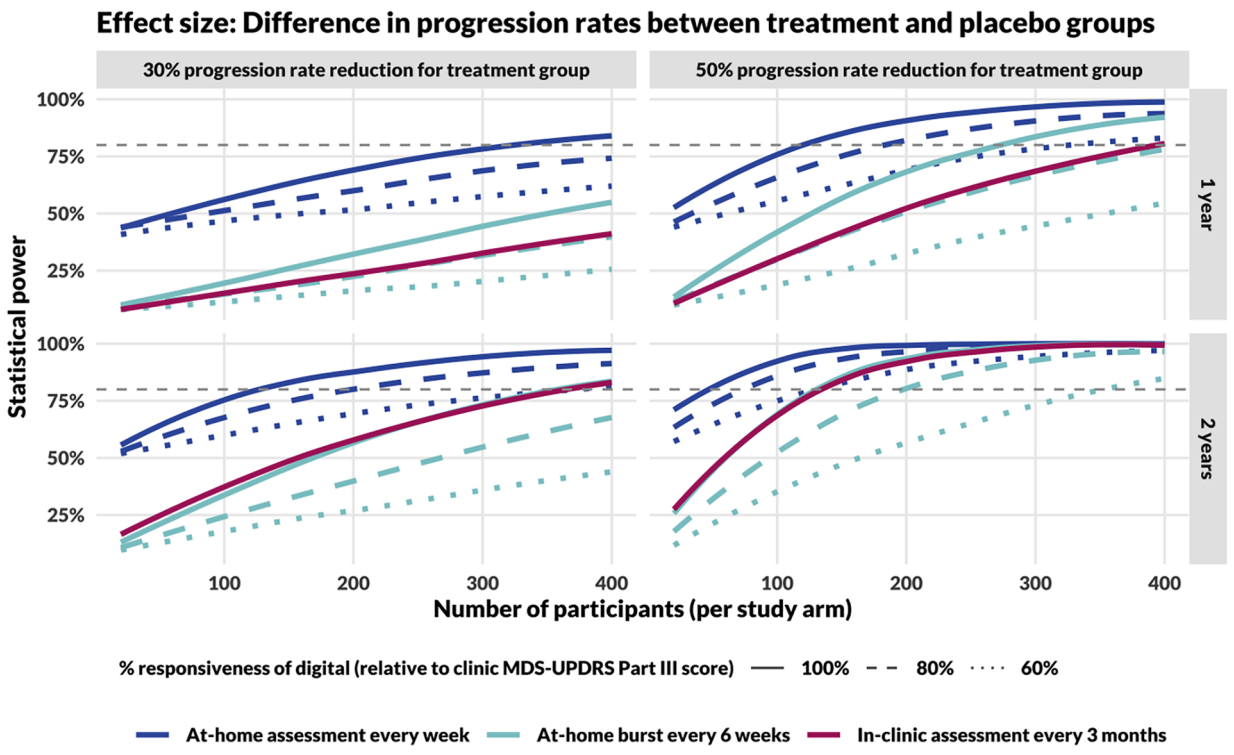


FIGURE 3
Power curves comparing study designs incorporating in-clinic and at-home assessments. The top row shows results for 1-year long studies, the bottom for 2-year. The left column models a DMT that reduces disease progression rate by 30% and the right column by 50%. In comparing DMT with placebo cohorts, the effect size for calculating study power is the difference in slopes of the measure over time (MDS-UPDRS Part III score for in-clinic assessments, Step Length during the 20-s walk test for at-home assessments), assessed using a linear mixed-effects model. Sample size calculations for in-clinic assessments (red, dashed line) assume responsiveness to progression and measurement error estimated by Evers et al. (15). The gray, dashed line represents the threshold for 80% power.

Additionally, the temporal spacing of at-home measures had a significant impact on statistical power. Study designs incorporating weekly assessments (48 assessments per year) consistently outperformed designs incorporating at-home bursts every 6 weeks (8 median bursts per year). Further, we found that a more even distribution of assessments always increased power under the assumption that the reliability was the same (Supplementary Figure 4). For example, 48 individual assessments provided greater power than 24 bursts of 2, which provided more power than 12 bursts of 4, and so on.

Based on an 80% statistical power threshold, we can make several different comparisons in sample size requirements between different study designs. For example:

- Assuming a 30% progression rate reduction and 100% responsiveness of the digital measure, a 2-year study would require approximately 110, 350, and 350 participants per study arm based on measures obtained weekly at-home, in 6-week bursts at home, and in-clinic every 3 months, respectively.
- Assuming a 50% progression rate reduction and 100% responsiveness of the digital measure, a 1-year study would require approximately 110, 270, and 390 participants per study arm based on endpoints obtained weekly at-home, in 6-week bursts at home, and in-clinic every 3 months, respectively.

We additionally considered the sensitivity of power calculations to estimates of trend and measurement error. The results indicated that in the presence of high variability in progression rates ($\sigma^2_T = 30$ for MDS-UPDRS part III total score), in the range estimated for PD (15), measurement error had little effect on statistical power (Supplementary Figure 5). In contrast, when progression rates had less variability (e.g., $\sigma^2_T = 1$ or 5), a more precise measure (e.g., $\sigma^2_m = 1$ or 5) substantially increased statistical power, especially for infrequent assessments. An increase in trend error of 20%–40% increases necessary sample sizes (Supplementary Figure 6), but its impact is less than that of a 20%–40% decrease in measure responsiveness (Figure 3).

3.3 Responsiveness of at-home measures

The responsiveness of digital measures to changes in motor function in PD is not yet well characterized; we therefore consider the impact of reduced responsiveness of a digital measure on the sample size needed for 80% power to detect a 30% reduction in progression rate in a treatment arm throughout a 1-year study (Figure 4). Using at-home assessments taken weekly would allow for detection of a modest 30% reduction in the rate of disease progression within

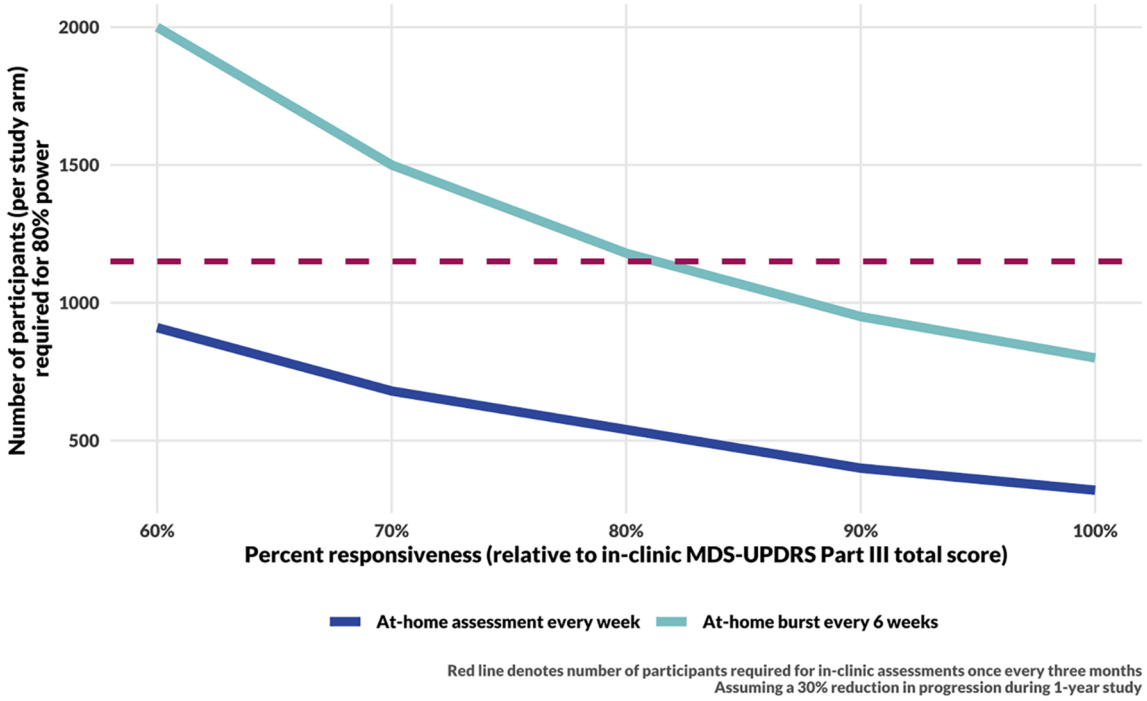


FIGURE 4
Sample size calculation results for a 1-year long DMT trial. The sample size for in-clinic assessments (red, dashed line) assumes responsiveness to progression and measurement error estimated by Evers et al. (15). Sample size calculations for DHT trials assume measurement error as estimated from the data (see Figure 2, Step Length) and consider a range of responsiveness of digital measures to underlying disease progression. The simulations for the blue curve include 48 assessments per year. The cyan curve includes 8 bursts per year, with 6 assessments per burst.

1 year with fewer than 910 participants per study arm even if the digital assessments were only 60% as responsive to progression as in-clinic MDS-UPDRS Part III total score. In the "ideal" scenario for which digital assessments are as responsive as-clinic MDS-UPDRS Part III, assuming a 30% reduction in the rate of disease progression, such a study would require 320 participants per arm compared to 1,150 per arm in a study that assesses MDS-UPDRS Part III in-clinic every 3 months.

3.4 False positive rate and minimum detectable change

Finally, we considered the implications of this modeling approach on the false positive rate. We found that while there was no strong evidence of bias in the estimates (Supplementary Figure 7), the probability of finding a significant difference between study arms when there was none (i.e., type I error) increased with both assessment frequency and trend variance (σ^2_T, Supplementary Figure 8). One way to manage this is to consider not only statistical but also clinical significance of the results. Indeed, the problem of type I error is mitigated if a simulation is considered to demonstrate study success if and only if the following two criteria are met: (1) the *p*-value for the difference in rates of change between treatment arms is <0.05 and (2) the estimated mean difference in the measure is greater than the minimum detectable change (MDC) (Supplementary Figure 9), though as expected, the probability of study success is reduced in this scenario.

4 Discussion

We estimated the reliability of a suite of at-home digital assessments administered on a smartphone to measure motor function in PD and performed simulations of clinical trial designs to assess the ramifications of implementing in-home digital health measures in DMT studies. In agreement with estimates of the reliability of other digital PD measures, we found the test-retest reliability for bursts of digital measures were as good as or better than the reliability of MDS-UPDRS part III scores. Interestingly, even though individual digital assessments typically have poorer test-retest reliability than in-clinic or at-home burst assessments, we found that a study design with evenly spaced digital weekly assessments outperformed both alternatives. This result suggests that the key challenge in measuring PD progression stems not from a lack of sufficiently sensitive and reliable measurement tools, but rather from the inherent variability in PD disease burden at points in time that renders infrequent measurement insufficient.

The result of superiority of frequent at-home assessments to in-clinic assessments every 3 months is robust to substantially decreased responsiveness of digital at-home measures compared with in-clinic (Figure 3). However, the quantitative results

regarding the necessary sample size were greatly affected by the responsiveness, and this will be important in future trial design. This is a difficult parameter to estimate as it requires longitudinal data. Ongoing and future multi-year studies that incorporate frequent digital measures in PD will be necessary to quantify this [e.g., (6, 27, 28)].

The results of this study suggest that evenly spaced assessments provide greater power than any configuration of an equal number of assessments distributed in bursts. This may be understood in the context of information theory; when compressing data using a logically irreversible process, such as summarizing a burst of assessments with a median, there is inherent loss of information as measured, for example, by Shannon entropy (29). The superiority of evenly spaced assessments also has implications for the implementation of DHTs in clinical trials. Frequent, evenly spaced measures require participants to consistently perform digital assessments across long periods of time. Adherence to at-home assessment regimens in clinical trials may decrease over time [e.g., (30)], and methods for maintaining usage will be important. Additionally, understanding the causes and impacts of missing assessments will be important.

We note that the results assumed progression rates and variability estimated in a patient population on standard of care medications such as levodopa (15). DMT studies are often longitudinal and conducted in patients in the early stages of PD [e.g., (31, 32)], which can include treatment-naive participants. Smaller sample sizes may be sufficient to detect DMT effects in treatment naive individuals, in part because the estimated progression rate is higher in the absence of medication (2). However, while participants may be unmedicated at the start of the study, over the course of a year or more they are likely to start symptomatic treatment (33). This transition can be challenging to account for in models of disease progression, and whether inclusion of covariates such as levodopa equivalent daily dose (LEDD) is sufficient to account for the changes induced by starting treatment remains an open question. There is substantial evidence that digital measures can detect levodopa effects [e.g., (20, 34, 35)], but as of yet, little evidence of detecting progression (14). Further work is needed to identify what clinical variables will be necessary to disentangle temporary fluctuations from underlying disease progression.

A key assumption in this study is that progression in both the treatment and placebo groups, while highly variable, is on average linear with time. Varying rates of progression with time could occur due to intrinsic characteristics of the motor function being measured, a learning effect, or time-dependent treatment effects of a DMT. Prior studies provide evidence for two of these: linear models in time are suitable for some but not all digital measures (14), and learning effects can be detected in at-home measures [e.g., (36, 37)]. As there are no approved DMTs for PD, the importance of time-dependent treatment effects remains unknown, but it is considered in other similar modeling assumptions (38) and is likely relevant. For measures whose progression cannot be approximated as linear, a study design that facilitates treating time as a discrete variable, such as bursts of assessments, may be beneficial. It should also be noted that this study does not model subpopulations within PD that may have different mean progression rates (16). Further work is necessary to understand how this type of heterogeneity in a population may affect the benefits and study design of digital at-home assessments. Additionally, data collected at higher frequency can require consideration of autocorrelation and temporal confounders (39).

One drawback of the mixed effects modeling approach taken in this study for power calculations is the possibility for false positive results. While estimates of trend using linear mixed effect models are largely insensitive to model misspecification (40), the standard error of the fixed effects may be underestimated in the presence of misspecified random effects such as autocorrelation (41, 42). The increased false positive rate with frequent sampling observed in the simulations can be understood in the context of the mismatch between the data generation process (i.e., a random walk with trend) and the model fitting procedure. As described here, one solution to this problem is to require not only statistical but also clinical significance. However, this comes with a loss of power to detect small changes, especially in shorter time windows. Analysis methods tailored to data that arise from underlying processes with this type of autocorrelation may be important in this context (43).

Future work to better understand the biological mechanisms underlying the progression of motor symptoms in PD can inform choices of models used for detecting treatment effects. In this study, the data generating process was chosen because it has been shown to parsimoniously explain heterogeneity present in the disease's dynamics in PPMI data (15), and therefore seems a reasonable candidate for a mechanistic model. The model assumes the mean underlying progression rate is the same across all patients, which we know to be an oversimplification. For example, certain genotypes progress more quickly than others [e.g., (44)]. One outcome of this assumption is that the trend variance reported by Evers (15) may be an overestimate as it accounts for not only random variation across time but also consistent variation between individuals that exists among the PPMI patients.

The mixed model framework used for effect detection has been used in longitudinal assessments of PD progression, including in PPMI data (2). However, while the model may appear to be a reasonable fit based on standard examination of residuals, our results suggest that care needs to be taken to avoid overconfidence in detection of small effects. Given the trade-off between power to detect treatment effect and the false positive rate that results from fitting misspecified models, future work to investigate the underlying mechanisms of motor function progression and the empirical autocorrelation structure of PD measurements will be important. Digital measures may provide an important window into the nuances of PD progression and its variability and allow for empirical examination of temporal correlation structures in data that can help determine optimal analysis methods (20). Given the high and increasing burden of PD around the globe, therapies that can stop or slow its progression will benefit millions of people (45). As of 2023, there were 63 ongoing clinical trials for PD DMTs, including 32 phase

II and 6 phase III (46). For these trials to be successful, in addition to an effective therapeutic agent, they must utilize measurements that allow for detection of treatment effect in the face of the high degree of variability inherent to PD progression. This study demonstrates that frequent measures enabled by digital health technologies that can be used consistently in patients' homes may increase the power to detect treatment effects in smaller and shorter trials.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: [doi: 10.7303/syn4993293].

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

JL: Conceptualization, Formal Analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. AS: Conceptualization, Formal Analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. SH: Formal Analysis, Writing – original draft, Writing – review & editing. JB: Writing – review & editing. EI: Conceptualization, Writing – review & editing. LO: Conceptualization, Data curation, Methodology, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was funded by Koneksa Health.

Acknowledgments

Data used for this paper were contributed by users of the Parkinson mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse [doi: 10.7303/syn4993293].

Conflict of interest

JL, AS, SH, JB, EI, and LO were employed by Koneksa Health.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2024.1430994/full#supplementary-material

References

1. Huang Y, Li Y, Pan H, Han L. Global, regional, and national burden of neurological disorders in 204 countries and territories worldwide. *J Glob Health*. (2023) 13:04160. doi: 10.7189/jogh.13.04160

2. Holden SK, Finseth T, Sillau SH, Berman BD. Progression of MDS-UPDRS scores over five years in de novo Parkinson disease from the Parkinson's progression markers initiative cohort. *Mov Disord Clin Pract*. (2018) 5:47–53. doi: 10.1002/mdc3.12553

3. Gonzalez-Robles C, Weil RS, van Wamelen D, Bartlett M, Burnell M, Clarke CS, et al. Outcome measures for disease-modifying trials in Parkinson's disease: consensus paper by the EJS ACT-PD multi-arm multi-stage trial initiative. *J Parkinsons Dis*. (2023) 13:1011–33. doi: 10.3233/JPD-230051

4. Burq M, Rainaldi E, Ho KC, Chen C, Bloem BR, Evers LJW, et al. Virtual exam for Parkinson's disease enables frequent and reliable remote measurements of motor function. *NPJ Digit Med*. (2022) 5:65. doi: 10.1038/s41746-022-00607-8

5. 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) 33:1287–97. doi: 10.1002/mds.27376

6. Lipsmeier F, Taylor KI, Postuma RB, Volkova-Volkmar E, Kilchenmann T, Mollenhauer B, et al. Reliability and validity of the Roche PD Mobile application for remote monitoring of early Parkinson's disease. *Sci Rep*. (2022) 12:12081. doi: 10.1038/s41598-022-15874-4

7. Oyama G, Burq M, Hatano T, Marks WJ Jr, Kapur R, Fernandez J, et al. Analytical and clinical validity of wearable, multi-sensor technology for assessment of motor function in patients with Parkinson's disease in Japan. *Sci Rep*. (2023) 13:3600. doi: 10.1038/s41598-023-29382-6

8. Martinez-Martin P, Rodriguez-Blazquez C, Alvarez-Sanchez M, Arakaki T, Bergareche-Yarza A, Chade A, et al. Expanded and independent validation of the movement disorder society-unified Parkinson's disease rating scale (MDS-UPDRS). *J Neurol*. (2013) 260: 228–36. doi: 10.1007/s00415-012-6624-1

9. 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) 23:2129–70. doi: 10.1002/mds.22340

10. Goetz CG, Stebbins GT. Assuring interrater reliability for the UPDRS motor section: utility of the UPDRS teaching tape. *Mov Disord*. (2004) 19:1453–6. doi: 10.1002/mds.20220

11. Williams S, Wong D, Alty JE, Relton SD. Parkinsonian hand or clinician's eye? Finger tap bradykinesia interrater reliability for 21 movement disorder experts. *J Parkinsons Dis*. (2023) 13:525–36. doi: 10.3233/JPD-223256

12. Roussos G, Herrero TR, Hill DL, Dowling AV, Muller MLTM, Evers LJW, et al. Identifying and characterising sources of variability in digital outcome measures in Parkinson's disease. *NPJ Digit Med*. (2022) 5:93. doi: 10.1038/s41746-022-00643-4

13. Jankovic J. Motor fluctuations and dyskinesias in Parkinson's disease: clinical manifestations. *Mov Disord*. (2005) 20(Suppl 11): S11–6. doi: 10.1002/mds.20458

14. Taylor K, Lipsmeier F, Scelsi M, Volkova-Volkmar E, Rukina D, Popp W, et al. Exploratory sensor-based outcome measures show divergent slopes of motor sign progression in Parkinson's disease patients treated with prasinezumab. *PREPRINT* (Version 1). (2024). doi: 10.21203/rs.3.rs-3921378/v1

15. Evers LJW, Krijthe JH, Meinders MJ, Bloem BR, Heskes TM. Measuring Parkinson's disease over time: the real-world within-subject reliability of the MDS-UPDRS. *Mov Disord*. (2019) 34:1480–87. doi: 10.1002/mds.27790

16. Fereshtehnejad S-M, Zeighami Y, Dagher A, Postuma RB. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. *Brain*. (2017) 140:1959–76. doi: 10.1093/brain/awx118

17. Marras C, Arbatti L, Hosamath A, Amara A, Anderson KE, Chahine LM. What patients say: large-scale analyses of replies to the Parkinson's disease patient report of problems (PD-PROP). *J Parkinson's Dis*. (2023) 13:757–67. doi: 10.3233/JPD-225083

18. PPMI: Parkinson Progression Marker Initiative. The Parkinson progression marker initiative (PPMI). *Prog Neurobiol*. (2011) 95:629–35. doi: 10.1016/j.pneurobio.2011.09.005

19. Ellis R, Kelly P, Huang C, Pearlmutter A, Izmailova ES. Sensor verification and analytical validation of algorithms to measure gait and balance and pronation/supination in healthy volunteers. *Sensors*. (2022) 22:6275. doi: 10.3390/s22166275

20. 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) 40:480–87. doi: 10.1038/s41587-021-00974-9

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. Van der Elst W, Molenberghs G, Hilgers RD, Verbeke G, Heussen N. Estimating the reliability of repeatedly measured endpoints based on linear mixed-effects models. A tutorial. *Pharm Stat*. (2016) 15:486–93. doi: 10.1002/pst.1787

23. Weir J. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. (2005) 19:231–40. doi: 10.1519/15184.1

24. Pinheiro J, Bates D, R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models. R Package Version 3*. (2023). p. 1–164. Available online at: https://cran.r-project.org/web/packages/nlme/index.html (Accessed September 24, 2024).

25. Team, R Core. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (2023).

26. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. (2016) 15:155–63. doi: 10.1016/j.jcm.2016.02.012

27. National Library of Medicine. *Usability, Validity and Biomarker Discovery for Wearable & Mobile Device Measurements of Neurologic Disorders*. (2024). Available online at: https://clinicaltrials.gov/study/NCT06219629 (Accessed September 24, 2024).

28. National Library of Medicine. *Assessing Tele-Health Outcomes in Multiyear Extensions of PD Trials (AT-HOME PD)*. (2018). Available online at: https://clinicaltrials.gov/study/NCT03538262 (Accessed September 24, 2024).

29. Hodge S, Vieland V. Information loss in binomial data due to data compression. *Entropy*. (2017) 19:75. doi: 10.3390/e19020075

30. Parra AG, Gonzalez-Medina G, Perez-Cabezas V, Casuso-Holgado MJ, Vinolo-Gil MJ, Garcia-Munoz C. Dropout rate of participants in randomized clinical trials that use virtual reality to train balance and gait in Parkinson's disease. A systematic review with meta-analysis and meta-regression. *J Med Syst*. (2023) 47:46. doi: 10.1007/s10916-023-01930-7

31. National Library of Medicine. *A 18-Month Study to Evaluate the Efficacy, Safety, Tolerability and Pharmacokinetics of Oral UCB0599 in Study Participants with Early-Stage Parkinson's Disease (ORCHESTRA)*. (2020). Available online at: https://clinicaltrials.gov/study/NCT04658186 (Accessed September 24, 2024).

32. National Library of Medicine. *A Study to Evaluate the Efficacy and Safety of Intravenous Prasinezumab in Participants with Early Parkinson's Disease (PADOVA)*. (2021). Available online at: https://clinicaltrials.gov/study/NCT04777331 (Accessed September 24, 2024).

33. Verschuur CVM, Suwijn SR, Boel JA, Post B, Bloem BR, van Hilten JJ, et al. Randomized delayed-start trial of levodopa in Parkinson's disease. *N Engl J Med*. (2019) 380:315–24. doi: 10.1056/NEJMoa1809983

34. Liu Y, Zhang G, Tarolli CG, Hristov R, Jensen-Roberts S, Waddell EM, et al. Monitoring gait at home with radio waves in Parkinson's disease: a marker of severity, progression, and medication response. *Sci Transl Med*. (2022) 14(663): eadc9669. doi: 10.1126/scitranslmed.adc9669

35. Griffiths RI, Kotschet K, Arfon S, Xu ZM, Johnson W, Drago J, et al. Automated assessment of bradykinesia and dyskinesia in Parkinson's disease. *J Parkinsons Dis*. (2012) 2:47–55. doi: 10.3233/JPD-2012-11071

36. 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:e26608. doi: 10.2196/26608

37. Czech MD, Badley D, Yang L, Shen J, Crouthamel M, Kangarloo T, et al. Improved measurement of disease progression in people living with early Parkinson's disease using digital health technologies. *Commun Med (Lond)*. (2024) 4:49. doi: 10.1038/s43856-024-00481-3

38. Mori H, Wiklund SJ, Zhang JY. Quantifying the benefits of digital biomarkers and technology-based study endpoints in clinical trials: project moneyball. *Digit Biomark*. (2022) 6:36–46. doi: 10.1159/000525255

39. Chaibub Neto E, Perumal TM, Pratap A, Tediarjo A, Bot BM, Mangravite L, et al. Disentangling personalized treatment effects from "time-of-the-day" confounding in mobile health studies. *PLoS One*. (2022) 17:e0271766. doi: 10.1371/journal.pone.0271766

40. Schielzeth H, Dingemanse NJ, Nakagawa S, Westneat DF, Allegue H, Teplitsky C, et al. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol Evol*. (2020) 11:1141–52. doi: 10.1111/2041-210X.13434

41. Matuschek H, Kliegl R, Vasishth S, Baayen H, Bates D. Balancing type I error and power in linear mixed models. *J Mem Lang*. (2017) 94:305–15. doi: 10.1016/j.jml.2017.01.001

42. LeBeau B, Song YA, Liu WC. Model misspecification and assumption violations with the linear mixed model: a meta-analysis. *SAGE Open*. (2018) 8. doi: 10.1177/2158244018820380

43. Staudenmayer J, Buonaccorsi JP. Measurement error in a random walk model with applications to population dynamics. *Biometrics*. (2006) 62:1178–89. doi: 10.1111/j.1541-0420.2006.00615.x

44. Sadaei HJ, Cordova-Palomera A, Lee J, Padmanabhan J, Chen S-F, Wineinger NE, et al. Genetically-informed prediction of short-term Parkinson's disease progression. *NPJ Parkinsons Dis*. (2022) 8:143. doi: do: 10.1038/s41531-022-00412-w

45. Chandler C, Folse H, Gal P, Chavan A, Proskorovsky I, Franco-Villalobos C, et al. Modeling long-term health and economic implications of new treatment strategies for Parkinson's disease: an individual patient simulation study. *J Mark Access Health Policy*. (2021) 9:1922163. doi: 10.1080/20016689.2021.1922163

46. McFarthing K, Buff S, Rafaloff G, Fiske B, Mursaleen L, Fuest R, et al. Parkinson's disease drug therapies in the clinical trial pipeline: 2023 update. *J Parkinsons Dis*. (2023) 13:427–39. doi: 10.3233/JPD-239901

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 23 December 2024
DOI 10.3389/fdgth.2024.1472691

OPEN ACCESS

EDITED BY
Graham Jones,
Tufts Medical Center, United States

REVIEWED BY
Brant Chapman,
Bergen Community College, United States
Tommaso Di Libero,
University of Cassino, Italy

*CORRESPONDENCE
Siegfried Hirczy
✉ siegfried.hirczy@va.gov

RECEIVED 29 July 2024
ACCEPTED 12 December 2024
PUBLISHED 23 December 2024

CITATION
Hirczy S, Zabetian C and Lin Y-H (2024) The current state of wearable device use in Parkinson's disease: a survey of individuals with Parkinson's.
Front. Digit. Health 6:1472691.
doi: 10.3389/fdgth.2024.1472691

The current state of wearable device use in Parkinson's disease: a survey of individuals with Parkinson's

Siegfried Hirczy[1,2]*, Cyrus Zabetian[1,2] and Yi-Han Lin[2]

[1]Department of Neurology, Veterans Affairs Puget Sound Health Care System, Seattle, WA, United States, [2]Department of Neurology, University of Washington, Seattle, WA, United States

Background: Interest in wearable device use in Parkinson's disease (PD) has grown rapidly with many compelling studies supporting diagnostic and therapeutic uses. Concurrently, consumer devices have proliferated and their role in health and wellness has expanded. However, incorporation of consumer and medical wearable devices into medical care has in our experience been limited.
Objective: We sought to assess the current state of consumer and medical wearable device use among those with PD and to understand the factors impacting their rate of use.
Methods: An anonymous online survey of individuals with PD in the US was conducted from July 9th, 2023, to Jan 8th, 2024, with 298 completed responses collected.
Results: Greater than 90% of respondents were interested in new technologies with 67% having had experiences with consumer wearable devices. Only 24% were using consumer devices for disease management and many functions were not fully utilized. Medical wearable device use was very limited with only 8% having used a device. Users of both consumer and medical wearables generally reported low barriers to use despite continued strong perceptions on the importance of cost, impact on care, comfort, and other factors.
Conclusion: This study demonstrates that for the clinical management of PD there is limited use of wearable devices even among individuals who are motivated and experienced with consumer wearable device use. Additionally, it is suggested that substantial barriers to medical wearable use are likely originating from the provider and/or systemic level.

KEYWORDS

Parkinson's disease, wearable, clinical practice, survey, barriers to use, consumer devices, medical devices

Introduction

Parkinson's disease (PD) is a neurodegenerative condition which results in abnormal movements, cognitive changes, and autonomic dysfunction. Diagnosis and management are made challenging due to inherent fluctuations in the severity of disease manifestations as well as dynamic changes that are induced by treatment. Standard clinical practice can only capture brief snapshots of the patient's experience and relies heavily on subjective report and quasi-objective exams, thus ability to optimally intervene is limited (1–4). Recently, there has been considerable interest in addressing these challenges through the creation of objective and continuous measures which hope

to allow for better understanding of each person's unique disease state and thereby improve treatment and reduce disability (5).

Wearable devices have been of great interest in healthcare due to their ability to contain imbedded sensors that help track various physiologic signals. Research in wearables in Parkinson's disease has been focused on the evaluation of motor features, and various devices have data supporting their ability to detect sub-clinical motor features, disease state fluctuations, disease progression, and to assist with therapeutic adjustments (6–9). However, various physiologic signals such as sleep (10), heart rate variability (11), cerebral oxygenation (12), and many others have been studied in PD with wearables. Studies in this area have rapidly increased over time with PubMed entries for "wearable" AND "Parkinson's Disease" going from a mere seven in 2012 to 170 in 2022. To help researchers and clinicians understand this landscape, many excellent reviews are available (13–18).

However, collection of valid data alone is not sufficient to change practice. The patient's perspective on device usability and utility is critical, and this has not been forgotten. Many studies of specific devices have included patient perspectives on features of interest, usability, and barriers to use (19–21). Additionally, more conceptually focused studies using surveys and focused groups have identified key features of interest such as wearability, ability to provide feedback, and clinical accuracy (22, 23). All this research has ultimately culminated in several medical wearable devices that are validated, designed with the patient in mind, and cleared for clinical use.

Concurrent with this explosion of research and approval of medical wearable use, there has been substantial adoption of consumer wearable devices for health tracking and lifestyle management. Devices such as the Apple Watch and Fitbit were in 2020 estimated to be used by around 25%–30% of the US population for health monitoring (24). These consumer devices also appear to have merits in PD as they offer the ability to potentially improve medication adherence, encourage and track physical activity, document symptoms, monitor sleep, and collect various other forms of information.

However, despite the extensive research on validation, the understanding of the factors important to patient users available to device manufacturers, the widespread use of consumer wearable devices, and the availability of approved medical devices for clinical use, real-world clinical data appears to be scant. In our clinical experience and after discussion with colleagues, few individuals are using these devices. While issues with the payor model for device use, lack of clinical impact, poor tolerability for patient and clinician users, and general disinterest in new technologies are commonly mentioned as factors playing a role, the evaluation of these barriers has not been extensively evaluated in routine clinical care.

We therefore sought to conduct a comprehensive evaluation of the current usage of wearable devices in PD, and to go beyond the controlled research setting to understand the real-world usage of wearable devices both consumer and medical. Additionally, we wanted to understand what factors were currently playing a role in current device usage and whether these were the same as those reported previously.

Methods

An anonymous online survey was conducted from July 9th, 2023, to Jan 8th, 2024. Respondents were self-identified individuals with PD and were requested to be at least 18 years of age.

Wearables were defined in this survey as any technological accessory which is affixed to the surface of an individual and which provides information on their movements (monitoring devices). A medical wearable device definition was not supplied, but options were explicitly listed (Apple Watch with StrivePD, PKG, KinesiaU, PDMonitor). Of note, given that StrivePD is an application that functions on a consumer device, we asked users of StrivePD to answer both as consumer wearable device Apple Watch users and as medical wearable device users.

The survey was designed by the study team with input from other specialist clinicians. Question topics included basic demographics, disease state, understanding and use of wearable devices (divided into consumer and medical device categories), general perceptions as related to theoretical devices, and general barriers to use. Survey questions were generally multiple-choice questions, but free response sections were provided in many cases to allow participants to provide answers that were not accounted for by the survey developers (Supplementary Survey Document).

Recruitment was conducted by collaborating groups who distributed study-related information and a link to the survey. The American Parkinson Disease Association (APDA) and the Washington State Parkinson Disease Registry participated in participant outreach (25).

Interested individuals followed the link and were brought to the online REDCap electronic data capture tool. The first page of the survey provided potential participants with information about the study and associated risks and benefits. Interested individuals would electronically confirm that they consented to participate, which would then allow them to proceed to the survey content (26, 27).

After survey completion, PD disease status nor any other characteristics of participants were verified. It was determined that verification would have limited the response rate and would have introduced more risk of identification and more bias as the systems available for use would tie individuals to specific medical systems. Additionally, it was believed that false representation was unlikely, as the survey was targeted, a response required substantial effort, and no notable financial incentive was present for respondents.

After survey closure, data processing and statistical analysis were performed using the R statistical analysis platform. Targeted sub-group analysis evaluating the effects of demographic and disease features on perceptions and experiences was performed post-hoc.

Results

Survey response, demographics, and disease state

A total of 346 responses were collected with 298 completed surveys (86%). The response rate was unknown but was

suspected to be very low given the size of the APDA distribution network. Only completed surveys were included in the analysis. Responses came from individuals living in 28 states with the greatest number coming from Washington State (63%) (Supplementary Table S1). A limited set of demographic features were recorded (Table 1). Disease related symptoms and characteristics varied encompassing both early and late stages of disease (Supplementary Tables S2, S3).

TABLE 1 Respondent demographic characteristics ($n = 298$).

Characteristics	Count	Percentage
Age		
<40	1	0.3%
40–50	9	3.0%
50–60	38	12.8%
60–70	110	36.9%
70–80	120	40.3%
80+	20	6.7%
Gender		
Male	128	43.0%
Female	169	56.7%
Prefer not to answer	1	0.3%
Residential setting		
Suburban	155	52.0%
Urban	92	30.9%
Rural	51	17.1%
Care setting		
Private/Non-university	180	60.4%
Tertiary/Academic	77	25.8%
VA/National Gov.	25	8.4%
County/Public	16	5.4%

Technology and consumer device experiences

Regarding technology and wearable device use, there was a high degree of interest in new technologies with 91% of individuals either very or somewhat interested. Knowledge about and use of wearable devices was also high with 87% knowing about wearable devices, 67% having used a device, and 56% currently using one (Figure 1). Most respondents knew about smart watches and fitness trackers (Supplementary Table S4); the Apple Watch was the most used and most preferred consumer device among respondents (Supplementary Table S5). Device use retention rates were also high with 84% of those with experience with wearable devices continuing to use a device. Among all device users, device usage time was very high with 90% using their preferred device nearly always or at least all the time while awake (Supplementary Table S6).

Consumer wearable device feature use

However, use of wearable devices for the management of PD was low at only 24% of respondents. Among those using wearable devices, individuals were using them most to track physical activity, medication timings, and sleep (Figure 2). Post-hoc subgroup analysis of Apple Watch wearable device users was performed to evaluate whether device capability limitations were playing a large role in lack of management related use, however findings mirrored those seen among all device users (Supplementary Figure S1).

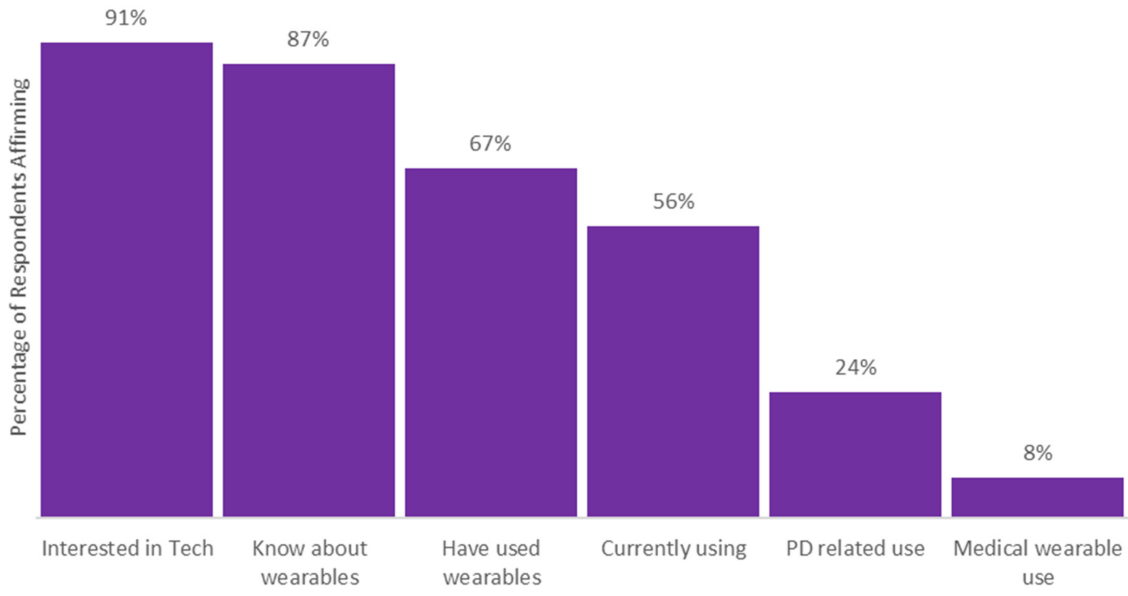


FIGURE 1
Respondent interest, knowledge, or experiences with wearable device use. For the interested in technologies category answers of moderately or very interested were combined, neutral or below were not incorporated. All other questions were yes or no responses.

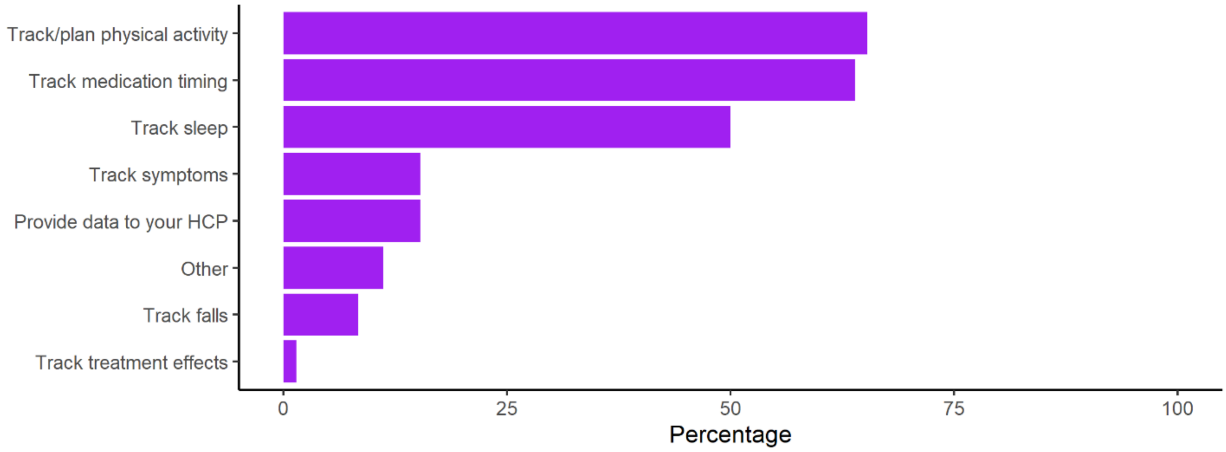


FIGURE 2
Use of specific consumer device functions for the management of PD. Calculated as the percentage of respondents using each function from the subset of respondents who affirmed that they were using consumer wearables to manage PD ($n = 72$).

Consumer wearable devices effect on disease management

Among those using wearables to manage PD, 76% reported a positive impact on personal management of PD (38% substantially, 39% somewhat), 24% reported that use was not particularly impactful, and no negative responses were recorded. Impact on medical team management of PD was felt to be not particularly impactful by 57% of respondents, though substantially positive (10%) or somewhat positive (33%) responses were reported, and no negative responses were reported.

Consumer wearable device barrier to use

Among survey respondents, current device use was limited most frequently because of lack of knowledge about the abilities of wearable devices and by cost (Supplementary Table S7). More generally, individuals were also surveyed on the use of smartphones and other applications for PD monitoring with 26% reporting the use of an application.

Medical wearable device usage and barrier to use

Medical wearable device use among respondents was 8% ($n = 23$). There were 19 StrivePD, 3 PD Monitor, 1 PKG, and 0 Kinesia users. Device use frequency was variable and many limitations on use were noted (Table 2). Impact on personal management was 26% positive (9% substantially, 17% somewhat) and impact on medical team management was 30% positive (4% substantially, 26% somewhat) (Supplementary Table S8).

General wearable interest and limitations

To separate perceptions and experiences tied to current consumer or medical devices, we asked about two similar theoretical devices. We described a version which provided information to the patient but did not directly provide it to the healthcare team (Type A) and one that provided information directly to both (Type B). For both versions, individuals were interested in using such a device (Supplementary Figure S2). Additionally, among those who would consider using the devices greater than 90% of respondents were willing to use either device at least all day while awake (Supplementary Figure S3). However, only 49% of those interested were willing to pay for a device if not covered by insurance. Of those who were willing to pay for such a device, the median one-time payment was $200 for both and the mean $252 (Type A) and $259 (Type B). Alternatively, we also asked about subscription pricing and what individuals would be willing to pay with the median being $10 for both devices and mean being $14 (Type A) and $15 dollars (Type B). There was no difference between the cost individuals were willing to pay for device [$p = 0.6$ (lump sum) and $p = 0.2$ (subscription), Wilcoxon paired signed rank test in the setting of non-normality of data shown by Shapiro-Wilk testing].

Finally, individuals were surveyed on barriers to their use of wearable devices for PD (Figure 3). Respondents reported concerns about all surveyed barriers with most respondents reporting at least a moderate level of concern for 4 out of the 5 surveyed barriers. Respondents reported the largest proportion of extreme concern for cost and impact on care. Additionally, we compared perceptions between individuals with different levels of wearable device experience (Supplementary Figure S4).

10.3389/fdgth.2024.1472691

TABLE 2 Respondent medical wearable device use, frequency of use, and limitations on use. Other limitations reported for Strive PD were that the device was in the process of being setup (1), difficulties logging in (1), geographic limitations on use (1), and an allergic response (1).

	Device			
	Any	StrivePD	PDMonitor	PKG
Total users	23	19	3	1
Use frequency				
Nearly always	43%	47%	33%	0%
All day when awake	26%	32%	0%	0%
>50% of wake time	4%	0%	33%	0%
Not daily but more than 3 days a week	4%	5%	0%	0%
Weekly	0%	0%	0%	0%
Monthly	0%	0%	0%	0%
Less than monthly	0%	0%	0%	0%
No longer using	22%	16%	33%	100%
Limitations on use				
No limitations	43%	47%	33%	0%
Uncomfortable/Difficult to keep on	0%	0%	0%	0%
Too much effort to maintain	9%	11%	0%	0%
Insufficient capabilities	9%	11%	0%	0%
Data input difficulties	17%	16%	0%	100%
Data review or access difficulties	4%	5%	0%	0%
Lack of impact on PD	26%	21%	67%	0%
Not utilized by healthcare provider	13%	11%	33%	0%
Concerns about accuracy of data	4%	5%	0%	0%
Discontinued by healthcare provider	4%	0%	33%	0%
Cost issues	0%	0%	0%	0%
Other	17%	21%	0%	0%

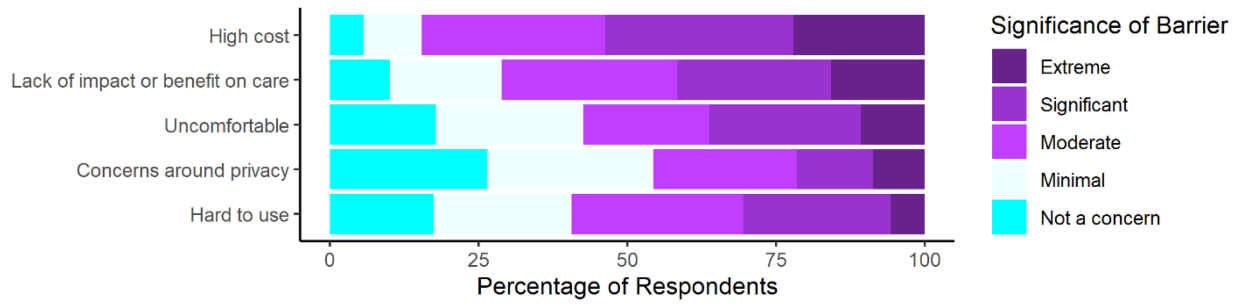


FIGURE 3
Perceived significance of certain barriers to respondents' personal use of wearable devices for management of PD as rated on a 5-point scale from not a concern to an extremely significant barrier (n = 298).

Discussion

This study captured wearable device experiences and perceptions among individuals with PD. Respondents were very technologically inclined (91% reporting interest) and were more likely to use wearable devices than the general US population, 67% vs. 35% (Morning Consult, Survey, 2023). Consumer devices among respondents also appeared to be well tolerated as discontinuation of wearables was rare and users also wore the devices most of the time.

However, despite the barriers to general consumer use being overcome, consumer wearable use for the specific purpose of managing PD was strikingly low at 24% and use of commonly available devices functions was low. This incomplete pattern of usage remained true even after removing ambiguity in the definition of "management" and ensuring that all functions queried were possible. In this subset, despite previous reported interest (28) and the importance placed on these features in PD, less than three quarters reported using the device to track physical activity, less than half tracked sleep, and less than a quarter tracked symptoms. Participants reported the lack of PD related use to be most often due to knowledge of functionality (26.2%). Features previously noted as important to address such as difficulties with wearability or comfort (4.7%), data input

(9.4%), and data review (5.0%) were not prominently reported (23). These findings are supportive of the acceptability and usability of current consumer devices in PD.

While consumer wearable device use for PD was sub-optimal, medical wearable device use was marginal. Only 8% of respondents had any experience with them, which was less than 1/8th the number of consumer wearable device users. However, there did not appear to be marked barriers to use once implemented, given 43% of respondents reported no limitations. The most noted barrier to use was lack of impact on care 26% and this was additionally supported by most medical device users indicating that their devices had negligible impact on their healthcare providers' management of PD and even on their own management. However, other issues appear to be reasonably addressed with less than 20% reporting difficulties with data input, less than 10% issues with wearability, and less than 5% issues with data review.

To better understand what factors were limiting wearable use and to compare current perceptions to prior research, many questions were directed to assess their perceptions on the significance of certain barriers, and it was again seen that individuals with PD were concerned about comfort, usability, and impact on care, as well as cost and privacy (23). However, these results seemingly conflict with the results obtained from direct questioning about their personal wearable use. Due to this conflict we sought to assess whether this was due to differences between users with more and less experience. However, consumer and medical wearable users still appeared to endorse similar levels of concern. These findings therefore suggest that while certain factors are still of high importance to people with PD they have generally been addressed by the current generation of devices.

Ultimately, it appears that the barriers to medical wearable device use and to a lesser degree consumer wearable device use in Parkinson's disease do not stem from individuals with PD. Even when an individual with PD is motivated, experienced with wearable use, and interested in theoretical medical wearable devices there is still a high likelihood that they will not be a wearable device user. We believe that this reflects difficulties with technology and device integration at the level of the provider user and/or the healthcare system.

The main strengths of this study were the number of respondents, the diversity of disease stages included, the inclusion of multiple wearables, and the level of detail obtained about perceptions and experiences. As with most survey studies, there were meaningful limitations. Foremost among them was sampling bias which was likely substantial given the online format and low estimated respondent rate. Respondents were likely highly motivated and technologically inclined. The use patterns and perceptions noted in this study do not directly reflect those of the population. However, they retain value as they are almost certainly a reflection of the upper bound of device use and their perceptions likely reflect the most positive reflections of the population, as such one can reasonably infer that the population rate of experience is lower and that perceptions are likely to be less positive than were seen in this sample. Additionally, the survey was heavily biased towards the Pacific Northwest region of the United States despite the goal of having a national distribution, demographic and socioeconomic data was limited, and the number of respondents for questions relating to consumer wearable device use in PD and medical wearables was relatively small.

Future studies should work to better understand the true perceptions of the PD population by expanding distribution, reducing respondent barriers, collecting more demographic and socioeconomic data, and engaging the community. Furthermore, future longitudinal studies should be performed to assess the evolution of individual perceptions of wearables as they evolve in relation to exposure to wearable devices and disease progression.

In conclusion, this study confirms the existence of a highly motivated subpopulation of individuals with PD who have a strong interest in wearable devices and confirms the feasibility of high levels of wearable device use in real-world use. Novelly, it identifies and partially quantifies large gaps in the use of consumer wearable device health tracking features and integration of wearable devices into PD related health management. Additionally, it confirms that medical wearable device use is low, but suggests that this isn't an issue with patient usability, thereby implicating provider and/or systemic barriers as the bottleneck to medical device use. We believe that these results call for further investigation into understanding the barriers affecting real-world use among clinician users and healthcare systems, as well as studies targeted at enhancing the utility and understanding of all forms of wearable device use in PD.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by University of Washington, Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin because human subjects could not be readily ascertained and data was not of a highly sensitive nature.

Author contributions

SH: Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Writing – original draft, Writing – review & editing. CZ: Conceptualization, Methodology, Resources, Supervision, Visualization, Writing – review & editing. YL: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Center for Advancing Translational Sciences, National Institutes of Health through a block grant (UL1 TR002319). Dr. Hirczy was supported by the Veterans Affairs Advanced Fellowship Program in Parkinson's Disease.

Acknowledgments

We wish to acknowledge our collaborators who made this survey possible, Dr. Rebecca Gilbert and the American Parkinson Disease Association (APDA) and the Washington State Parkinson Disease Registry. We also wish to acknowledge Dr. Alberto Espay for his assistance with the development of the survey.

Conflict of interest

YL reports a relationship with Boston Scientific Corporation, CereGate GmBH, and Bial, all of which provide funding grants.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2024.1472691/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Use of specific consumer device functions for the management of PD. Calculated as the percentage of respondents using each function from the subset of respondents who affirmed that they were using consumer wearables to manage PD and who were using the Apple Watch as their preferred device ($n = 41$).

SUPPLEMENTARY FIGURE S2
Respondent degree of interest in using theoretical wearable devices in PD. Device A was defined as a device with the ability to help individuals better understand their disease and responses to medical interventions, but without ability to directly provide information to healthcare providers. Device B was similarly defined but with the added benefit of healthcare provider integration ($n = 192$).

SUPPLEMENTARY FIGURE S3
Amount of time respondents were willing to wear theoretical devices A and B if there were no issues with comfort or usability. (Device A $n = 293$, Device B $n = 292$).

SUPPLEMENTARY FIGURE S4
Significance of barriers to wearable device use among individuals with different degrees of experience with wearable device use. MWU = medical wearable device user, CWU + PD = consumer wearable device user who manages PD with the device. CWU-PD = consumer wearable device user not managing PD with the device. Ex CWU = participants who have used consumer wearable devices but no longer are. Non-User = no experience with wearable devices ($n = 298$).

SUPPLEMENTARY TABLE S1
States in which survey respondents reside ($n = 298$).

SUPPLEMENTARY TABLE S2
Respondent PD related characteristics and symptom characteristics ($n = 298$).

SUPPLEMENTARY TABLE S3
Respondent assistive gait device utilization frequencies among those who reported using an assistive device ($n = 58$).

SUPPLEMENTARY TABLE S4
Respondent awareness of and experiences with wearable devices ($n = 298$).

SUPPLEMENTARY TABLE S5
Respondent wearable use preferences among those with experience with wearable devices ($n = 193$). Other responses for preferred device: apollo neuro (1), iphone (1), kospet (1), and Samsung watch (1).

SUPPLEMENTARY TABLE S6
Preferred consumer wearable device use time among individuals still using thei preferred device ($n = 160$).

SUPPLEMENTARY TABLE S7
Respondent reported current barriers to the use of consumer wearable devices for the management of PD. Individual respondents could affirm the presence of multiple barriers.

SUPPLEMENTARY TABLE S8
Medical wearable device user perceptions on impact of devices on management of PD ($n = 23$).

References

1. Evers LJW, Krijthe JH, Meinders MJ, Bloem BR, Heskes TM. Measuring Parkinson's disease over time: the real-world within-subject reliability of the mds-updrs. *Mov Disord.* (2019) 34(10):1480–7. doi: 10.1002/mds.27790

2. Regnault A, Boroojerdi B, Meunier J, Bani M, Morel T, Cano S. Does the Mds-updrs provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. *J Neurol.* (2019) 266(8):1927–36. doi: 10.1007/s00415-019-09348-3

3. Papapetropoulos SS. Patient diaries as a clinical endpoint in Parkinson's disease clinical trials. *CNS Neurosci Ther.* (2012) 18(5):380–7. doi: 10.1111/j.1755-5949.2011.00253.x

4. Shulman LM, Pretzer-Aboff I, Anderson KE, Stevenson R, Vaughan CG, Gruber-Baldini AL, et al. Subjective report versus objective measurement of activities of daily living in Parkinson's disease. *Mov Disord.* (2006) 21(6):794–9. doi: 10.1002/mds.20803

5. Ymeri G, Salvi D, Olsson CM, Wassenburg MV, Tsanas A, Svenningsson P. Quantifying Parkinson's disease severity using mobile wearable devices and machine learning: the parkapp pilot study protocol. *BMJ Open.* (2023) 13(12):e077766. doi: 10.1136/bmjopen-2023-077766

6. Farzanehfar P, Woodrow H, Horne M. Sensor measurements can characterize fluctuations and wearing off in Parkinson's disease and guide therapy to improve motor, non-motor and quality of life scores. *Front Aging Neurosci.* (2022) 14:852992. doi: 10.3389/fnagi.2022.852992

7. Schalkamp A-K, Peall KJ, Harrison NA, Sandor C. Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis. *Nat Med*. (2023) 29(8):2048–56. doi: 10.1038/s41591-023-02440-2

8. Powers R, Etezadi-Amoli M, Arnold EM, Kianian S, Mance I, Gibiansky M, et al. Smartwatch inertial sensors continuously monitor real-world motor fluctuations in Parkinson's disease. *Sci Transl Med*. (2021) 13(579):eabd7865. doi: 10.1126/scitranslmed.abd7865

9. Sotirakis C, Su Z, Brzezicki MA, Conway N, Tarassenko L, FitzGerald JJ, et al. Identification of motor progression in Parkinson's disease using wearable sensors and machine learning. *NPJ Parkinson's Dis*. (2023) 9(1):142. doi: 10.1038/s41531-023-00581-2

10. Ko YF, Kuo PH, Wang CF, Chen YJ, Chuang PC, Li SZ, et al. Quantification analysis of sleep based on smartwatch sensors for Parkinson's disease. *Biosensors*. (2022) 12(2):74. doi: 10.3390/bios12020074

11. Suzuki M, Nakamura T, Hirayama M, Ueda M, Hatanaka M, Harada Y, et al. Wearable sensor device-based detection of decreased heart rate variability in Parkinson's disease. *J Neural Transm*. (2022) 129(10):1299–306. doi: 10.1007/s00702-022-02528-y

12. Joo JY, Yoo D, Kim JM, Shin C, Ahn TB. Effect of positional changes on cerebral perfusion in Parkinson's disease patients with orthostatic hypotension. *J Mov Disord*. (2024) 17(4):408–15. doi: 10.14802/jmd.24104

13. Moreau C, Rouaud T, Grabli D, Benatru I, Remy P, Marques A-R, et al. Overview on wearable sensors for the management of Parkinson's disease. *NPJ Parkinson's Dis*. (2023) 9(1):153. doi: 10.1038/s41531-023-00585-y

14. Mughal H, Javed AR, Rizwan M, Almadhor AS, Kryvinska N. Parkinson's disease management via wearable sensors: a systematic review. *IEEE Access*. (2022) 10:35219–37. doi: 10.1109/ACCESS.2022.3162844

15. Ancona S, Faraci FD, Khatab E, Fiorillo L, Gnarra O, Nef T, et al. Wearables in the home-based assessment of abnormal movements in Parkinson's disease: a systematic review of the literature. *J Neurol*. (2022) 269(1):100–10. doi: 10.1007/s00415-020-10350-3

16. Deb R, An S, Bhat G, Shill H, Ogras UY. A systematic survey of research trends in technology usage for Parkinson's disease. *Sensors*. (2022) 22(15):5491. doi: 10.3390/s22155491

17. Li P, van Wezel R, He F, Zhao Y, Wang Y. The role of wrist-worn technology in the management of Parkinson's disease in daily life: a narrative review. *Front Neuroinform*. (2023) 17:1135300. doi: 10.3389/fninf.2023.1135300

18. Sigcha L, Borzì L, Amato F, Rechichi I, Ramos-Romero C, Cárdenas A, et al. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: a systematic review. *Expert Syst Appl*. (2023) 229:120541. doi: 10.1016/j.eswa.2023.120541

19. Botros A, Schutz N, Camenzind M, Urwyler P, Bolliger D, Vanbellingen T, et al. Long-term home-monitoring sensor technology in patients with Parkinson's disease-acceptance and adherence. *Sensors*. (2019) 19(23):5169. doi: 10.3390/s19235169

20. Zhao Y, Heida T, van Wegen EE, Bloem BR, van Wezel RJ. E-Health support in people with Parkinson's disease with smart glasses: a survey of user requirements and expectations in The Netherlands. *J Parkinsons Dis*. (2015) 5(2):369–78. doi: 10.3233/JPD-150568

21. van den Bergh R, Evers LJW, de Vries NM, Silva de Lima AL, Bloem BR, Valenti G, et al. Usability and utility of a remote monitoring system to support physiotherapy for people with Parkinson's disease. *Front Neurol*. (2023) 14:1251395. doi: 10.3389/fneur.2023.1251395

22. Ozanne A, Johansson D, Hallgren Graneheim U, Malmgren K, Bergquist F, Alt Murphy M. Wearables in epilepsy and Parkinson's disease-a focus group study. *Acta Neurol Scand*. (2018) 137(2):188–94. doi: 10.1111/ane.12798

23. Kenny L, Moore K, O'Riordan C, Fox S, Barton J, Tedesco S, et al. The views and needs of people with Parkinson disease regarding wearable devices for disease monitoring: mixed methods exploration. *JMIR Form Res*. (2022) 6(1):e27418. doi: 10.2196/27418

24. Dhingra LS, Aminorroaya A, Oikonomou EK, Nargesi AA, Wilson FP, Krumholz HM, et al. Use of wearable devices in individuals with or at risk for cardiovascular disease in the us, 2019 to 2020. *JAMA Network Open*. (2023) 6(6):e2316634-e. doi: 10.1001/jamanetworkopen.2023.16634

25. Kim HM, Leverenz JB, Burdick DJ, Srivatsal S, Pate J, Hu S-C, et al. Diagnostic validation for participants in the Washington state Parkinson disease registry. *Parkinson's Disease*. (2018) 2018:3719578. doi: 10.1155/2018/3719578

26. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The redcap consortium: building an international community of software platform partners. *J Biomed Inform*. (2019) 95:103208. doi: 10.1016/j.jbi.2019.103208

27. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (redcap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. (2009) 42(2):377–81. doi: 10.1016/j.jbi.2008.08.010

28. Serrano JA, Larsen F, Isaacs T, Matthews H, Duffen J, Riggare S, et al. Participatory design in Parkinson's research with focus on the symptomatic domains to be measured. *J Parkinsons Dis*. (2015) 5(1):187–96. doi: 10.3233/JPD-140472

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 10 January 2025
DOI 10.3389/fdgth.2024.1500811

OPEN ACCESS

EDITED BY
Graham Jones,
Tufts Medical Center, United States

REVIEWED BY
Francesco Picariello,
University of Sannio, Italy
Joshua Chang,
The University of Texas at Austin, United States

*CORRESPONDENCE
Jie Ren
jie.ren@merck.com
Richard Baumgartner
richard_baumgartner@merck.com

RECEIVED 24 September 2024
ACCEPTED 16 December 2024
PUBLISHED 10 January 2025

CITATION
Zhai S, Liaw A, Shen J, Xu Y, Svetnik V, FitzGerald JJ, Antoniades CA, Holder D, Dockendorf MF, Ren J and Baumgartner R (2025) A novel machine learning based framework for developing composite digital biomarkers of disease progression. Front. Digit. Health 6:1500811. doi: 10.3389/fdgth.2024.1500811

A novel machine learning based framework for developing composite digital biomarkers of disease progression

Song Zhai[1], Andy Liaw[1], Judong Shen[1], Yuting Xu[1], Vladimir Svetnik[1], James J. FitzGerald[2,3], Chrystalina A. Antoniades[2], Dan Holder[1], Marissa F. Dockendorf[4], Jie Ren[4]* and Richard Baumgartner[1]*

[1]Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, NJ, United States, [2]NeuroMetrology Lab, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom, [3]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, United Kingdom, [4]Digital Clinical Measures, Translational Medicine, Merck & Co., Inc., Rahway, NJ, United States

Background: Current methods of measuring disease progression of neurodegenerative disorders, including Parkinson's disease (PD), largely rely on composite clinical rating scales, which are prone to subjective biases and lack the sensitivity to detect progression signals in a timely manner. Digital health technology (DHT)-derived measures offer potential solutions to provide objective, precise, and sensitive measures that address these limitations. However, the complexity of DHT datasets and the potential to derive numerous digital features that were not previously possible to measure pose challenges, including in selection of the most important digital features and construction of composite digital biomarkers.
Methods: We present a comprehensive machine learning based framework to construct composite digital biomarkers for progression tracking. This framework consists of a marginal (univariate) digital feature screening, a univariate association test, digital feature selection, and subsequent construction of composite (multivariate) digital disease progression biomarkers using Penalized Generalized Estimating Equations (PGEE). As an illustrative example, we applied this framework to data collected from a PD longitudinal observational study. The data consisted of Opal™ sensor-based movement measurements and MDS-UPDRS Part III scores collected at 3-month intervals for 2 years in 30 PD and 10 healthy control participants.
Results: In our illustrative example, 77 out of 235 digital features from the study passed univariate feature screening, with 11 features selected by PGEE to include in construction of the composite digital measure. Compared to MDS-UPDRS Part III, the composite digital measure exhibited a smoother and more significant increasing trend over time in PD groups with less variability, indicating improved ability for tracking disease progression. This composite digital measure also demonstrated the ability to classify between *de novo* PD and healthy control groups.
Conclusion: Measures from DHTs show promise in tracking neurodegenerative disease progression with increased sensitivity and reduced variability as compared to traditional clinical scores. Herein, we present a novel framework and methodology to construct composite digital measure of disease progression from high-dimensional DHT datasets, which may have utility in accelerating the development and application of composite digital biomarkers in drug development.

KEYWORDS

composite digital biomarker, Parkinson's disease, disease progression, linear mixed effects model, machine learning, penalized generalized estimating equations

1 Introduction

Neurodegenerative diseases, including Parkinson's Disease (PD), are an area of vast unmet medical need. Drug development efforts in this area have increasingly focused on the search for disease-modifying therapies that slow down the underlying disease progression mechanisms. However, a lack of validated measures that allow for disease progression to be monitored objectively, relatively rapidly, and with high precision makes it challenging to effectively demonstrate therapeutic efficacy and hinders drug development efforts. PD clinical trials generally use the Movement Disorder Society—Unified Parkinson's Disease Rating Scale (MDS-UPDRS) to track disease progression longitudinally. However, MDS-UPDRS is subjective in nature, relies on patient and caregiver-reported symptoms and clinician's qualitative ratings (1), is slow to change, and has low measurement precision, resulting in large and lengthy clinical trials to test efficacy for potential disease modifying therapies (2).

Recent advances in digital health technologies (DHTs) offer unprecedented opportunities to collect more objective, precise, and sensitive measures, both in the clinic and remotely, that were out of reach in the past. Such measures could provide new insights into neurogenerative disease progression, including for Parkinson's disease. There are many studies that have investigated using measures from sensor-based digital health technologies in neurodegenerative diseases (3–11). These studies have collectively demonstrated that many neurodegenerative disease symptoms can be quantified by DHTs. Moreover, multiple longitudinal observational studies have shown that digital measures can pick up changes over time that are indicative of disease progression (12–18). It is further thought that the objective measures enabled by DHTs could offer improved sensitivity and reduced variability (12, 19), which could translate to smaller and shorter clinical trial designs (20) and, in turn, potential for accelerated drug development. Despite promising results, the longitudinal studies published to date have used different DHTs and analysis methodologies to identify the digital features of importance and to derive respective digital clinical measures, making it difficult to compare across studies or create consensus among the research community. Open discussions on the methodology of digital clinical measure development and evaluation are critically needed to move the field forward.

It has been increasingly recognized that composite digital measures, rather than reliance on individual digital features, are needed for more effective measurement of disease progression as compared to traditional clinical composite scores. Adams et al. (21) showed that no individual digital feature (from gait, tremor, turns, speech, and cognition) outperformed MDS-UPDRS Part III (a composite clinical score) in terms of the standardized change from baseline after 12 months in a PD observational study (WATCH-PD). Furthermore, Czech et al. (22) demonstrated individual sensor-based digital features of upper and lower extremity bradykinesia often lacked strong sensitivity to longitudinal changes, whereas digital composite scores showed significant differences over 12 months in WATCH-PD.

There have been several examples where composite digital measures were developed for disease classification and/or tracking symptom progression (22–30); however, the approach taken has varied, and there have been limited discussions on the methodologies to effectively select informative digital features and construct the most performant composite measures. For example, Perumal and Sankar (23) developed a Linear Discriminant Analysis (LDA) classifier using multiple gait features collected from wearable sensors to distinguish between PD patients and healthy control (HC) subjects. Czech et al. (22) constructed composite digital scores using pre-defined combinations of features from single tasks (pronation-supination and toe-tapping) and used them to measure longitudinal progression of bradykinesia after 1 year. Sotirakis et al. (30) developed a Random Forest model to estimate the MDS-UPDRS III values using gait and sway features and used the model to detect progression of motor symptoms longitudinally. These efforts vary in terms of the measure construction (pre-defined vs. supervised ML, choice of models), the clinical label selection (MDS-UPDRS III total score or single item), the selection of digital tasks (single task e.g., toe-tapping or a combination of tasks), as well as the selection of input features (e.g., whether features are pre-screened). Overall, the field has not adopted consistent and systematic methods and/or analysis frameworks. Therefore, there is an urgent need to develop methodologies and analysis pipelines for the construction of composite digital measures for disease progression tracking, tailored for high-dimensional, longitudinal data with digital features collected from sensor technologies.

The types of data generated by DHTs are often longitudinal and high dimensional, which differs from traditional clinical measures, calling for novel analytical strategies to handle such data for the construction of composite digital measures. Unlike traditional clinical measures that collect a defined set of measures at each time point, DHTs leverage various sensors to generate large amounts of time-series data (e.g., acceleration, screen touch, audio/video, keyboard press), either collected from defined active task-based assessments or from passive monitoring. Such data are often not readily analysable statistically and need to be aggregated and transformed into digital features first. For example, for measurement of physical activity, continuous accelerometer signals are often converted to epoch level activity counts and then aggregated over time into features such as daily total activity count, total steps, non-sedentary time, etc., for further statistical analysis. There can be large numbers of features derived from the high-frequency sensor signals; such features may have various data types (i.e., categorical, continuous, duration, etc.) and clinimetric properties, many of which may not yet have been fully explored as it was not previously possible to measure them without use of DHTs. These features could have intrinsic skewness in distribution, floor/ceiling effects, as well as unknown redundancies and covariances. In addition, the high frequency nature of DHT data collection and potential for remote data acquisition can also lend itself to higher levels of data missingness. Furthermore, not all digital features that can be generated from sensor data may have clinical significance or be valuable for creating composite digital measures. These attributes of DHT data make it a unique

challenge in the development of composite digital measures to track longitudinal disease progression.

Machine learning (ML) methods offer a valuable tool for selecting the most informative digital features to reflect disease progression and to construct clinically meaningful composite digital measures. ML-based techniques can often improve prediction performance in analysing digital data in neurodegenerative diseases; however, existing ML methodologies for longitudinal data analysis are also challenged by the high dimensionality of DHT data. For example, although the generalized estimating equations (GEE) method (31) incorporating different patterns of working correlation matrix across multiple timepoints has been widely used in longitudinal data analysis, the direct use of classical unpenalized GEE in high-dimensional longitudinal data analysis may lead to misleading results (32). To address this, an ML-based penalized GEE (PGEE) method (32) could be used to improve upon the GEE method in handling DHT data. PGEE performs simultaneous coefficient estimation and variable selection for longitudinal data analysis with high-dimensional covariates by including a penalty term in the GEE model, which can be better-suited to handle high-dimensional feature sets.

In this paper, we propose a principled, scalable, and comprehensive methodology framework for the development of novel composite digital biomarkers, derived from DHT data and anchored to the MDS-UPDRS score, to measure neurodegenerative disease progression. This framework includes data processing, univariate digital feature screening, multivariate (composite) digital biomarker construction (using PGEE methods), and composite biomarker performance evaluation.

We further demonstrate the utility of this framework by applying it to a sample dataset containing high-dimensional, longitudinal movement data collected by a body-worn accelerometer system from a PD longitudinal observation study. The current analytical challenges of high-dimensional and longitudinal digital data and path forward for the application of composite digital biomarkers in measurement of neurodegenerative disease progression are also discussed.

2 Materials and methods

2.1 Study overview

To illustrate our proposed methodology to construct composite digital measures for tracking longitudinal disease progression, we applied the framework to data from 30 PD patients (10 *de novo* PD patients, 10 mild-to-moderate PD patients on levodopa, and 10 advanced PD patients) and 10 healthy control subjects from a PD longitudinal observational study conducted at John Radcliffe Hospital in Oxford, UK (11, 30, 33). The participants visited the clinic once every 3 months for 2 years. At each visit, they wore six synchronized inertial measurement units (IMUs) ("Opal" sensors, APDM Wearable Technologies, a Clario Company) across their body and performed two-minute walk, postural sway, and timed up-and-go (TUG) tasks. The Mobility Lab^{TM} software (APDM Wearable Technologies, a Clario Company) was then used to process these raw sensor signals and generate epoch-level digital features at each instance of a time period or physical movement (e.g., per minute, per step, per turn, or per sit-to-stand event). The MDS-UPDRS Part III assessments were also conducted at these clinic visits. The MDS-UPDRS Part III score and subscales (including Bradykinesia, Postural and Gait, Rigidity, and Tremor, defined in Supplementary Table S1) were calculated. Demographic data including age and sex of the participants were also collected at the beginning of the study.

2.2 Statistical analysis

The workflow of our proposed comprehensive machine learning based framework is illustrated in Figure 1, which comprises five main steps: (1) data collection and processing; (2) univariate feature screening; (3) univariate association testing; (4) multivariate analysis (using PGEE) to construct a composite digital measure for longitudinal disease progression; (5) performance evaluation. The specifics of each step are described below.

2.2.1 Data processing and quality control

In this first step, data aggregation and pre-processing are performed to convert high-frequency, epoch-level data into a set of aggregated digital features for each task. The movement data collected from DHTs often include epoch-level features (e.g., per second, per minute, or per walking step) that are collected repeatedly during an active task (e.g., two-minute walk). This step simplifies such data and produces a clean, high-dimensional feature set for each participant at each clinical time point, in order to facilitate subsequent longitudinal analyses.

In our illustrative PD example, summary statistics (mean, median, standard deviation, and mean absolution deviation) were calculated to represent the repeated measurements across the entire task for features that had repeated measurements during the task. For example, during the two-minute walk task, step lengths of every step that the participant took were recorded; these were aggregated into task-level features such as mean step length during the two-minute walk task period. After that, we had 256 digital features generated in total. Then, distributions of all features were examined, and the non-informative features that had few distinct values, included a large amount of data missingness, or contained extreme values were removed. For the remaining features, missing data imputation was performed using the mean of available data in each feature. Finally, additional feature quality control steps were implemented, which included removing highly correlated features, log-transforming skewed features, and removing outliers. 141 unique digital features were left for univariate progression screening in the next step.

2.2.2 Univariate progression screening

In the second step of our framework, univariate progression screening is recommended to identify whether each digital feature detected disease progression during the study duration. In this step, a linear mixed effects model (LMM) is used to screen the univariate features against a set of pre-determined criteria. Each

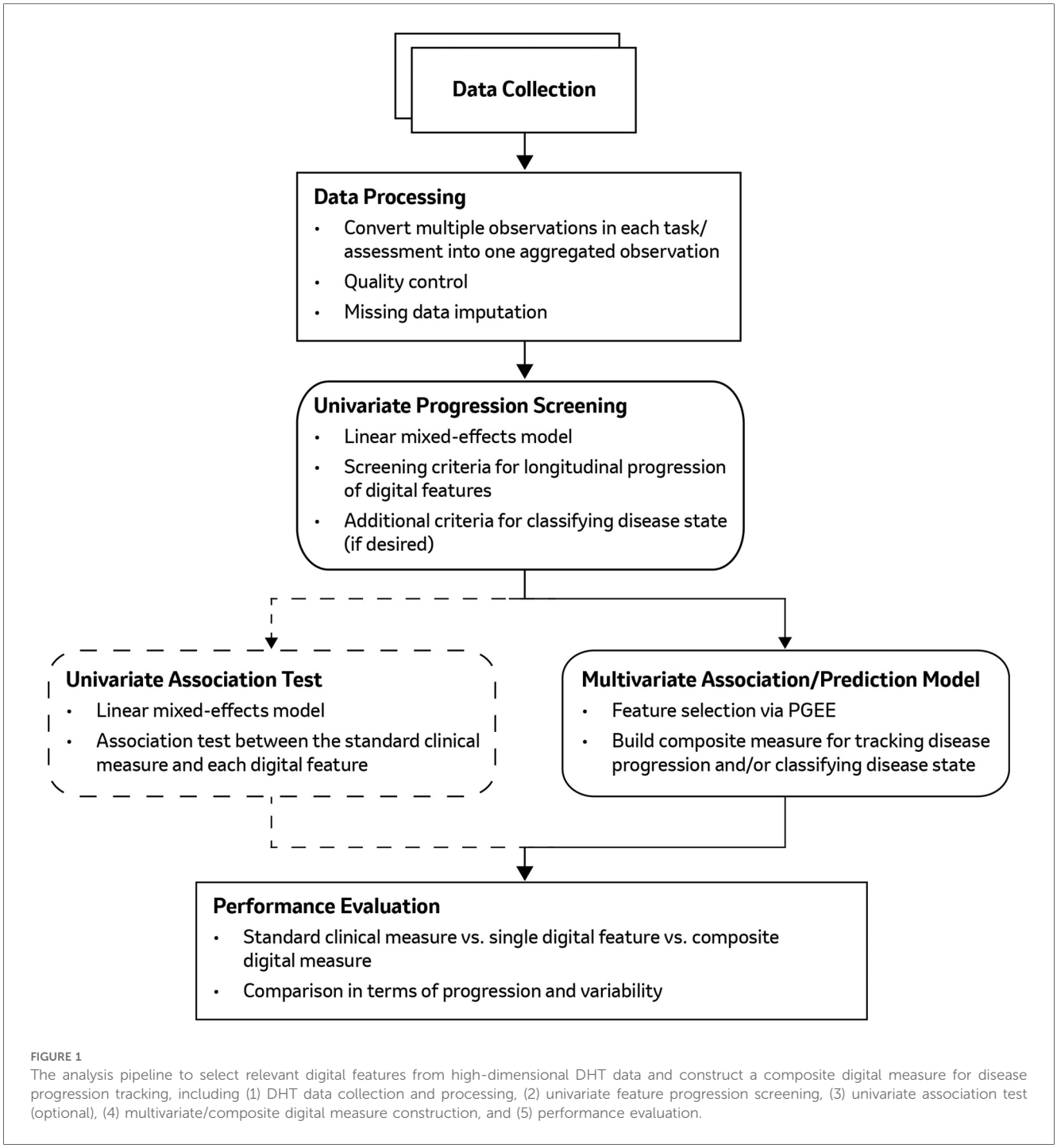


FIGURE 1
The analysis pipeline to select relevant digital features from high-dimensional DHT data and construct a composite digital measure for disease progression tracking, including (1) DHT data collection and processing, (2) univariate feature progression screening, (3) univariate association test (optional), (4) multivariate/composite digital measure construction, and (5) performance evaluation.

digital feature is used as the response variable for the screening separately. Independent variables are added to the model as fixed effects, including covariates to be adjusted, group membership, visit, group-by-visit interaction, and covariate-by-visit interactions. Random intercept and slope are added to the model as random effects.

In our illustrative PD example, we applied relatively relaxed screening criteria to select digital features for downstream analysis. We considered a digital feature as a "candidate" if (1) its longitudinal trend was flat in the HC group (i.e., the LMM slope p-value of HC group was larger than 0.05) and (2) it demonstrated a progression trend with time in PD groups (i.e., the LMM group-by-visit interaction p-value was <0.1 or the p-value of the differential slope between *de novo*/mild-to-moderate/advanced PD and HC was <0.1).

2.2.3 Univariate association test

To gain additional insights on the univariate associations between the standard clinical measure (i.e., MDS-UPRDS Part III) and the candidate digital features that passed the univariate progression screening, our framework employs a univariate association test step. In this step, a linear mixed effects model is

employed, with the clinical measure as the dependent variable and each individual digital feature as the independent variable. Covariates to be adjusted are also included in the model. Random intercepts for each subject are allowed in the model and *p*-values are calculated to assess the significance of the association between the clinical measure and digital features.

An optional procedure is to further filter the candidate digital features based on their associations to the standard reference measure (i.e., MDS-UPDRS Part III and its subscales in our example) and exclude non-significant features. In our example, we chose to implement relatively relaxed screening criteria to retain more features for the subsequent feature selection, and therefore, we did not exclude features that did not show association with MDS-UPDRS Part III in our downstream analysis.

2.2.4 Multivariate prediction model

In the final step of our framework, a multivariate prediction model is developed to select a subset of digital features from those that passed the univariate progression screening and combine them into a composite digital biomarker of disease progression.

For feature selection, we used the longitudinal MDS-UPDRS Part III data as the training endpoint in our illustrative example. Additionally, we included features that were important for classifying the *de novo* PD cohort from healthy controls in the feature selection process since patient identification could also be an important attribute for the composite digital biomarker. Importantly, depending on the intended context-of-use of the developed measure, one could use our proposed framework to optimize the measure for disease progression tracking, or patient identification, or both, by adjusting the screening criteria and the training endpoints used.

To model the high-dimensional longitudinal data, our framework includes a ML-based Penalized Generalized Estimating Equations (PGEE) method (32), which performs simultaneous coefficient estimation and variable selection. Compared to the traditional GEE method, PGEE introduces a penalty term to the estimating function of GEE (details of PGEE is provided in Supplementary Method S1).

To determine the optimal number of digital features (P) to be included into the final multivariate prediction model, a cross-validation (CV) strategy is implemented into the framework to avoid overfitting (34). Specifically, all digital features are first ranked by their PGEE estimates from the training set, then a series of PGEE models with different numbers of top features are built and evaluated in the testing set. The optimal number of features is then determined to be the number of features from the model with the smallest Root Mean Squared Error (RMSE). The approach is further described in Supplementary Method S2.

Once P is determined, the PGEE estimates of the digital features that pass the univariate screening are calculated again using the whole dataset, and the top P features with the largest PGEE estimates were selected. Two sets of digital features were identified based on a PGEE model for MDS-UPDRS Part III progression and another PGEE model for *de novo* PD vs. Healthy Control classification, respectively. These two feature sets were merged into a final feature set for the composite digital measure construction.

A GEE model is then fitted, with this final feature set plus the covariates as independent variables, and MDS-UPDRS Part III as a continuous dependent variable. This generates our final composite digital measure for performance evaluations.

3 Results

3.1 Patient demographics and baseline characteristics

The baseline demographic characteristics for the participants included in our illustrative analysis are shown in Table 1 and Supplementary Figure S1. The mean ages of four groups (*de novo* PD, mild-to-moderate PD, advanced PD, and HC) were 66.2, 61.6, 71.2, and 65.6 years, respectively. The ratios of male-to-female subjects in the four groups were 5:5, 9:1, 5:5, and 3:7, respectively.

To determine if age and sex needed be considered covariates to be adjusted for in our models, we calculated the age-by-visit and sex-by-visit interaction *p*-values in linear mixed effects models with MDS-UPDRS Part III as the response in the pooled PD group. The results, summarized in Supplementary Table S2, suggested that age would affect the slope of MDS-UPDRS Part III progression (with p-value = 0.04) while sex would not (with p-value = 0.19). We therefore considered only age as a covariate to be adjusted in our data analysis models.

3.2 Univariate progression screening results

In our illustrative example, our univariate progression screening criteria were such that a digital feature would "pass" if the LMM model for that digital feature showed (1) no progression in the control group and (2) a progression in at least one of the three PD groups. 77 digital features out of 141 screened passed these criteria, including 15 features from postural sway task, 5 features from timed up-and-go (TUG) task, and 57 features from two-minute walk task. Among these, Walk GLLGS (Gait—Lower Limb—Gait Speed) had the smallest group-by-visit interaction p-value (6.0×10^{-7}) and the smallest *de novo* PD vs. HC progression slope p-value (4.7×10^{-4}); Walk GLLDS

TABLE 1 Patient baseline characteristics (age and sex) for the three PD groups and healthy control group.

	de novo PD	Mild-to-moderate PD (on-therapy)	Advanced PD	Healthy Control
N	10	10	10	10
Age (years) [mean (SD)]	66.2 (6.46)	61.6 (10.76)	71.2 (4.78)	65.6 (6.98)
Sex [Male (%)/Female (%)]	5 (50)/5 (50)	9 (90)/1 (10)	5 (50)/5 (50)	3 (30)/7 (70)

(Gait—Lower Limb—Double Support) had the smallest mild-to-moderate PD vs. HC progression slope p-value (0.01); and Walk GLLSD (Gait—Lower Limb—Step Duration) had the smallest advanced PD vs. HC progression p-value (1.2×10^{-6}). P-values of TUG TPV (Timed Up and Go—Turn Peak Velocity) for group-by-visit interaction, *de novo* PD vs. HC progression slope, mild-to-moderate PD vs. HC progression slope, and advanced PD vs. HC progression slope were 0.008, 0.001, 0.147, and 0.013, respectively. A summary heatmap of all 77 digital features that met the screening criteria is displayed in Figure 2, and the heatmap of all the digital features that were screened is displayed Supplementary Table S3.

3.3 Univariate association analysis results

Figure 3 shows the univariate association testing results between the 77 digital features that passed the univariate screening in our illustrative example and MDS-UPDRS Part III scores (and its subscales). 37 of these 77 digital features (48.1%) showed significant associations (i.e., p-value < 0.05) with MDS-UPDRS Part III scores (including 32 features from the Walk task, 3 features from the TUG task, and 2 features from the Sway task). The associations of digital features with the MDS-UPDRS Part III scores were generally consistent with their associations with the Bradykinesia (BK) subscale within MDS-UPDRS Part III. Specifically, 40 of the 77 digital features were associated with the BK subscale (including 31 features from the Walk task, 3 features from the TUG task, and 6 features from the Sway task). In addition, 59 of the 77 digital features were associated with the Postural Instability and Gait (PIGD) subscale (including 54 features from the Walk task, 4 features from the TUG task, and 1 feature from the Sway task), while only 3 of the 77 features (TUG TPV, TUG TA, and Walk GULMV) were associated with the Tremor Dominant (TD) subscale.

Turn Peak Velocity (TPV), obtained from the Timed Up and Go (TUG) test (35), demonstrated the most significant association with MDS-UPDRS Part III. TUG TPV is defined as the maximum achieved angular velocity of trunk rotation in the y-axis during 180-degree turns (deg/sec) and has been found to be related to PD progression in multiple studies (12, 36–38). The progression characteristics of TUG TPV are shown in Figure 4, where the group-wise and subject-wise lines were obtained from the linear mixed effect model and the points represented the observed data. In terms of TUG TPV, the mild-to-moderate, on therapy PD and HC groups were stable, while the *de novo* and advanced PD groups showed progression.

In general, the univariate association observations were consistent with the progression patterns seen in the MDS-UPDRS Part III and its subscales, which is shown in Supplementary Figure S2. Specifically, compared to the HC group, the BK subscale progressed across all PD groups (at $\alpha = 0.1$ level). The PIGD subscale progressed in *de novo* and advanced PD groups while staying stable in the mild-to-moderate, on-therapy PD group. This pattern was similar to most of the digital features included in the analysis, as indicated in Figure 2. In contrast, the TD subscale progressed in the mild-to-moderate, on-therapy PD group, while remaining unchanged in *de novo* and advanced PD groups.

3.4 Multivariate feature selection and prediction results

3.4.1 Feature selection

We first conducted multivariate feature selection to determine the optimal number of features to be selected for inclusion into the composite score and prediction model in our illustrative example analysis. Supplementary Figure S3 indicated that for developing a composite digital measure for disease progression tracking, using 9 top features (ranked by their PGEE estimates in training sets during cross-validation) overall yielded the smallest RMSE; and for classifying disease status, using 3 top features resulted in the largest AUC via internal cross-validation.

We then ranked all pre-screened features (i.e., digital features) according to their PGEE estimates in the whole dataset supervised by the continuous endpoint (i.e., MDS-UPDRS Part III) and the binary endpoint (i.e., *de novo* PD vs. HC), respectively. Nine digital features (TUG TD, TUG TPV, TUG STSD, Walk TA, Walk GLLC, Walk GLLSW, Walk GLLLSM, Walk APAMAA, and Sway PSAN95ESA) were selected for disease progression tracking; additionally, three digital features (TUG TPV, Walk GLLTOA, and Walk GULMV) were selected for PD vs. Control classification. Table 2 lists the description of these selected features. The two sets of digital features were further merged; since one of the features (TUG TPV) was in both feature sets, 11 unique digital features were included in the final feature set. This feature set was then used to create the composite digital biomarker by fitting a GEE model.

3.4.2 Composite digital biomarker for tracking MDS-UPDRS part III

The performance of the composite digital biomarker was evaluated using 10-fold cross-validation in PD and HC groups, respectively. As shown in Figure 5, the composite digital measure showed a pattern of no change vs. time in the HC group as expected (with RMSE in HC group = 2.8). On the other hand, it had a smoother increasing trend in the overall PD group, as well as each PD subgroup (with RMSE in PD group = 12.7).

We further compared performances among MDS-UPDRS Part III, the composite digital measure, and each of the univariate digital features included in the composite digital measure (e.g., TUG TPV) quantitatively in terms of both progression and variability. Detailed results are summarized in Table 3. Overall, the group-by-visit interaction p-value of the composite digital measure was close to that of MDS-UPDRS Part III (7.65×10^{-3} vs. 6.22×10^{-3}). The increasing trend of the composite digital measure was much more significant compared to MDS-UPDRS Part III and individual digital feature TUG TPV in *de novo* and advanced PD groups. Specifically, for *de novo* PD vs. HC, the effect sizes of progression slope were 1.41, 2.14, and 1.37 for MDS-UPDRS Part III, composite digital measure, and TUG TPV,

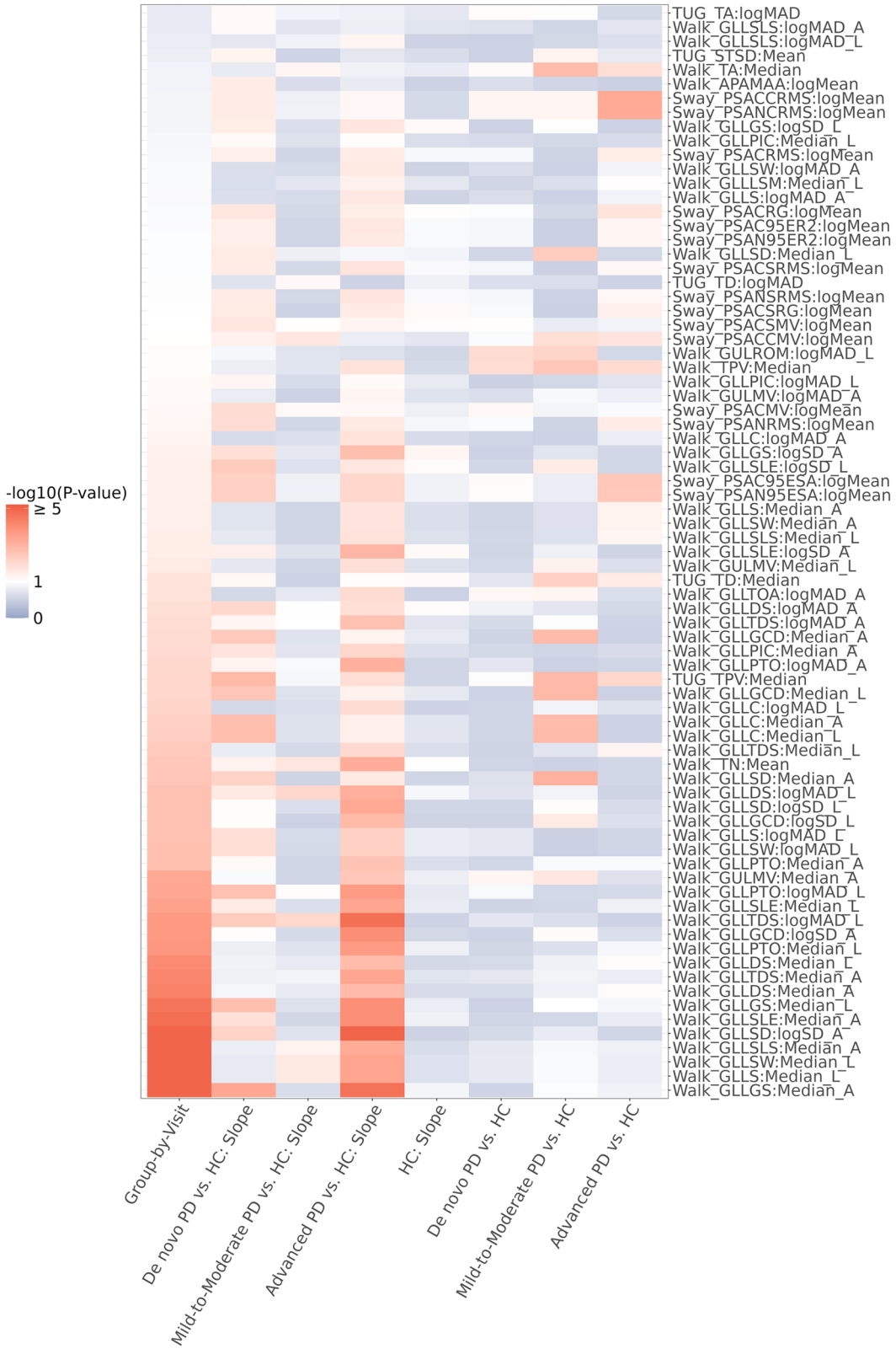


FIGURE 2
Heatmap representation of the *p*-values of the 77 digital features that passed the progression screening. The screening criteria applied were (1) no time progression in the HC group (i.e., LMM slope *p*-value of HC > 0.05), and (2) time progression in at least one of the three PD groups (i.e., LMM group-by-visit interaction *p*-value < 0.1 or *p*-value of differential slope between *de novo*/mild-to-moderate/advanced PD and HC < 0.1).

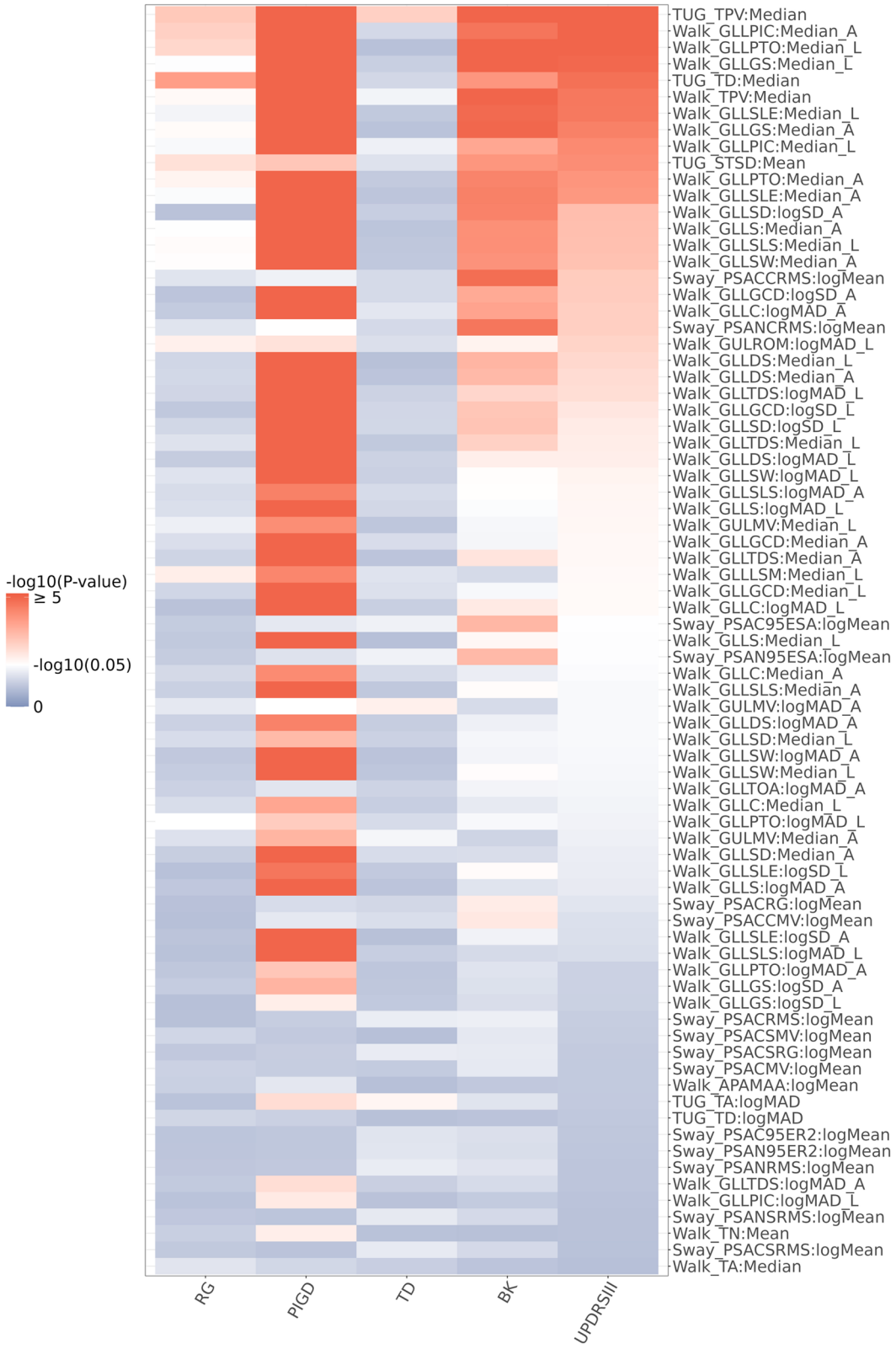


FIGURE 3
Heatmap of the univariate association testing *p*-values between MDS-UPDRS part III (and its subscales: BK, TD, PIGD, RG) and the 77 digital features that passed the univariate screening. *P*-values were calculated from a linear mixed effects model with MDS-UPDRS Part III or its subscales as the outcome variable. The 77 features were ranked based on their association *p*-values from the analysis with the MDS-UPDRS Part III score. Each digital feature and age were included as independent variables. Random intercept was added as a random effect.

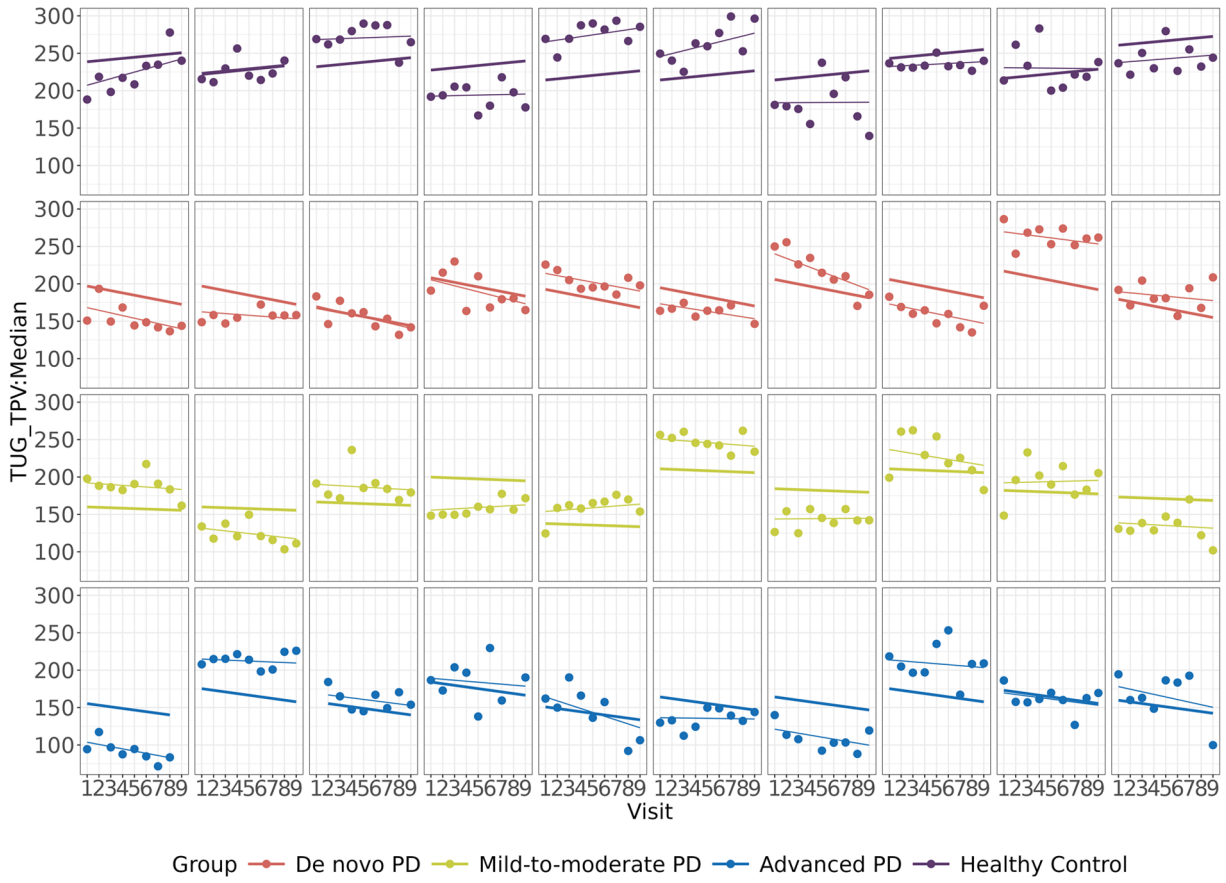


FIGURE 4
Results from a digital feature, TUG TPV: turn peak velocity (TPV), obtained from the timed Up and Go (TUG) test, which showed the most significant association with the MDS-UPDRS part III score. Each row represents the three PD groups and the HC group. Each panel within a row corresponds to a particular subject. The thick lines and thin lines denote the group-wise and subject-wise estimates of progression lines fitted by the linear mixed effects model, respectively. The points denote the observed data.

TABLE 2 Description of the selected features: 9 features selected for longitudinally disease progression tracking, and 3 features selected for *de novo* PD vs. HC classification.

Objective for feature selection	Feature	Statistic	Side	Description	PGEE Estimate
Disease progression tracking	TUG TD	Median		Turns—Duration	0.39
	TUG TPV	Median		Turns—Turn Velocity	−0.38
	TUG STSD	Mean		Stand to Sit—Duration	0.34
	Walk TA	Median		Turns—Angle	0.26
	Walk GLLC	MAD	A	Gait/Lower Limb—Cadence	0.16
	Walk GLLSW	MAD	L	Gait/Lower Limb—Swing	0.07
	Walk GLLLSM	Median	L	Gait/Lower Limb—Circumduction	−0.06
	Walk APAMAA	Mean		Anticipatory Postural Adjustment—Forward APA Peak	−0.06
	Sway PSAN95ESA	Mean		Postural Sway/Angles—Sway Area	0.05
de novo PD vs. HC classification	TUG TPV	Median		Turns—Turn Velocity	−0.60
	Walk GLLTOA	MAD	A	Gait/Lower Limb—Toe Out Angle	−0.43
	Walk GULMV	Median	A	Gait/Upper Limb—Arm Swing Velocity	−0.27

MAD, Mean Absolute Deviation; A, Affected side; L, Less affected side.

respectively. For advanced PD vs. HC, the effect sizes of progression slope were 0.76, 0.90, and 0.76 for MDS-UPDRS Part III, composite digital measure, and TUG TPV, respectively. On the other hand, the composite digital measure didn't show significant progression in the mild-to-moderate, on-therapy PD group, which was consistent with what is observed in Figure 5. Recall that none of the 11 selected digital features had significant univariate progression in the mild-to-moderate, on-therapy PD group (for example, the mild-to-moderate PD vs. HC slope p-value of TUG TPV was not significant, $p = 0.15$). Thus, it was not surprising that the

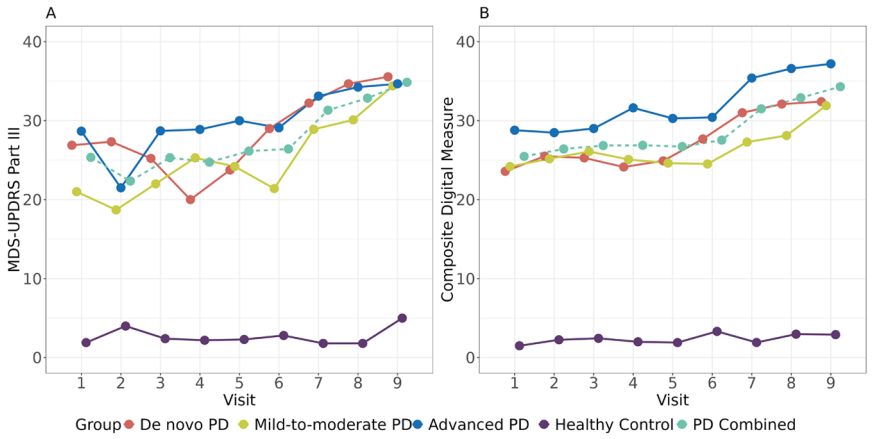


FIGURE 5
Tracking PD progression via (A) the MDS-UPDRS part III and (B) the composite digital measure based on the 11 selected digital features. The dashed lines represented the observed MDS-UPDRS Part III scores and the predicted composite digital measures in the combined PD group, respectively.

TABLE 3 Performance comparison among MDS-UPDRS part III, the composite digital measure, and TUG TPV in terms of both progression and variability.

	MDS-UPDRS Part III	Composite Digital Measure	TUG TPV
Group-by-visit *p*-value	6.22×10^{-3}	7.65×10^{-3}	8.05×10^{-3}
de novo PD vs. HC: slope *p*-value	0.02	8.28×10^{-3}	1.35×10^{-3}
Mild-to-moderate PD vs. HC: slope *p*-value	2.25×10^{-4}	0.16	0.15
Advanced PD vs. HC: slope *p*-value	0.07	4.22×10^{-3}	0.01
HC slope *p*-value	0.01	0.58	0.78
Between-subject coefficient of variation	39.0%	20.4%	17.2%
Within-subject coefficient of variation	34.1%	16.9%	9.7%
Effect size in progression slope between *de novo* PD and HC	1.41 (0.51, 2.31)	2.14 (1.24, 3.04)	1.37 (0.47, 2.27)
Effect size in progression slope between mild-to-moderate PD and HC	2.28 (1.40, 3.16)	0.86 (−0.02, 1.73)	0.60 (−0.28, 1.47)
Effect size in progression slope between advanced PD and HC	0.76 (−0.11, 1.64)	0.90 (0.02, 1.77)	0.76 (−0.12, 1.64)

composite digital measure preserved the same pattern. Moreover, the composite digital measure showed smaller between-/within-subject coefficient of variation than MDS-UPDRS Part III. In summary, the results from Figure 5 and Table 3 indicate that the composite digital measure is an attractive aggregated measure for tracking PD progression compared to MDS-UPDRS Part III and to individual digital features.

3.4.3 Performance in classifying *de novo* PD and HC

We further examined if the composite digital measure developed above (for tracking PD progression longitudinally) was also effective in classifying between *de novo* PD and HC subjects. Results are shown in Figure 6, where the boxplot of composite digital measures in the *de novo* PD subgroup is clearly higher than the boxplot in the HC subgroup. The composite digital measure had an AUC of 0.992 in such classification, which was very similar to that achieved for the classification model based on MDS-UPDRS Part III (AUC of 0.991). This demonstrated that the composite digital measure was able to preserve the ability to differentiate PD from HC groups and was effective in classifying *de novo* PD and HC.

4 Discussion

DHT-derived measures have shown great promise in both tracking disease progression and disease classification. However, it remains challenging to identify digital features for predicting disease progression longitudinally in a high dimensional space. Furthermore, methodologies for combining individual digital features into composite digital measures have not been fully explored and standardized in the field of DHTs. Although there have been several examples where composite digital measures were developed for tracking symptom progression, many of these prior efforts used simple sums of pre-defined, unweighted features without optimizing for performance (22, 29). In other cases where digital composite measures were trained/optimized to predict clinical scores, machine learning methodologies were often used without consideration of the longitudinal nature of the features (15, 30). The field has not adopted consistent and systematic methods and/or analysis frameworks that use statistical or machine learning methods capable of handling high-dimensional longitudinal data for feature selection and prediction while considering the within-subject correlation across visits. Therefore, there is an urgent need to develop this kind of new methodologies

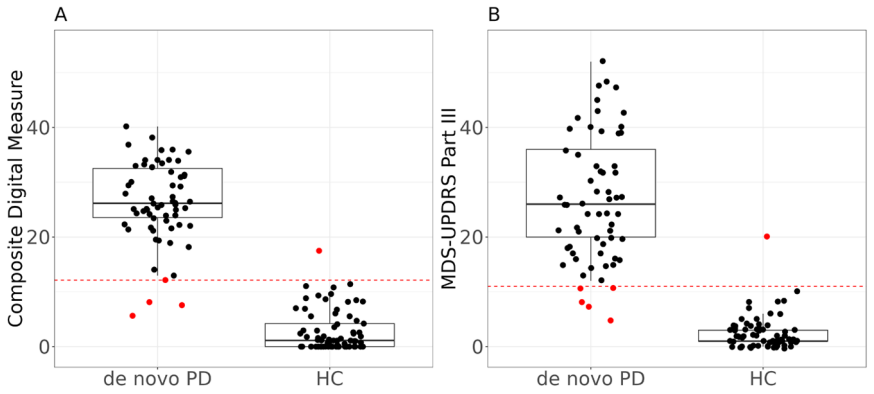


FIGURE 6
Boxplots of (**A**) the composite digital measure vs. (**B**) the MDS-UPDRS Part III in *de novo* PD and HC groups, respectively. Each dot indicates the composite digital measure or clinical score of each subject at each visit; the dotted line indicates the optimized threshold for classification: 12.14 for composite digital measure and 11 for MDS-UPDRS Part III. Dots above the line were classified into *de novo* PD, and dots below the line were classified into HC. Values that lead to false classifications are shown in red. The composite digital measure has AUC of 0.992 and MDS-UPDRS Part III has AUC of 0.991.

and analysis pipelines for the construction of composite digital measures for disease progression tracking, tailored for high-dimensional, longitudinal data with digital features collected from sensor technologies. In this paper, we propose a principled, scalable, and comprehensive methodology for the identification of relevant digital features of disease progression from large DHT data sets, and subsequent construction of a composite digital measure for disease progression tracking. Specifically, in Step 1, data is collected and processed for aggregated observation and quality control. In Step 2, we apply a linear mixed effects model for univariate screening for longitudinal progression of digital features. In Step 3, a univariate association test is conducted between candidate digital features (i.e., features that pass the univariate screening) and clinical scores, for example the MDS-UPDRS Part III and/or its subscales. In Step 4, the candidate digital features are ranked via a ML-based method, PGEE, for high-dimensional longitudinal data analysis. The optimal number of top features to be included into the composite digital measure is further determined using a cross-validation based algorithm to avoid overfitting. Note that PGEE (penalized GEE) method is designed for longitudinal data analysis with high-dimensional covariates by including a penalty term in the traditional GEE model. PGEE is particularly useful in handling high-dimensional feature sets, applicable for data from DHTs.

To demonstrate the utility of our methodology, we applied it to the data collected from a PD longitudinal observational study, which consisted of Opal^{TM} sensor-based movement measurements and MDS-UPDRS Part III scores collected from PD patients at a range of disease stages and healthy controls over a 2-year duration. Our primary interest in developing a composite digital measure is to track disease progression. The composite digital measure developed from this illustrative example generally showed a smoother and more significant increasing trend in PD groups and smaller between-/within-subject coefficients of variation than MDS-UPDRS Part III in this small dataset ($N = 40$), indicating potential utility for the composite digital measures to be used to track disease progression more sensitively and with less variability vs. standard clinical measures. It should be noted that the dataset in our illustrative example was small ($N = 40$), and therefore, results of our analysis should be interpreted with caution. The analysis reported here was presented as an illustration of our proposed methodology and framework and was not intended as a proposed composite measure for use in future studies. We also note that the composite digital measure shows less significant progression trending in mild-to-moderate, on-therapy PD patients compared to in *de novo* and advanced PD patients. This outcome is consistent with the trends observed by Brzezicki et al. (11) using data derived from the OxQUIP study. We further evaluated the classification performance between *de novo* PD and HC using the composite digital measure built from our methodology (primarily for tracking PD progression). The measure had an AUC ROC of 0.992 for classification (vs. AUC = 0.991 when using MDS-UPDRS Part III), indicating that the composite digital measure also had a good performance in classifying between *de novo* PD and HC subjects, comparable to MDS-UPDRS Part III.

Note that in our analysis, the top digital features (i.e., those with the largest PGEE estimates from the multivariate penalized regression model) were selected for both tracking MDS-UPDRS Part III progression and classifying between *de novo* PD and HC. While the digital feature TUG TPV ranked high in both subsets of selected features, we observe that the digital features that are important for disease progression tracking are not necessarily the same as digital features important for patient identification. We constructed the composite digital biomarker with the merged feature list in this example.

Regarding performance, this composite digital biomarker keeps the main characteristics of individual digital features but exhibits a more significant increasing trend indicative of disease progression.

On one hand, this composite digital biomarker shows progression in *de novo* and advanced PD groups but remains flat in mild-to-moderate, on-therapy PD and HC groups which is consistent with the general behaviours of individual digital features (Figure 2). On the other hand, it exhibits a more significant increasing longitudinal trend compared to individual digital features (including TUG TPV). It is worth noting that the features selected using the PGEE model tend to have diverse measurement properties. The final feature set is not a combination of the best-performing individual features in terms of individual progression signal (i.e., neither features with the most progressions in PD groups nor features with the most significant univariate association with MDS-UPDRS Part III). A possible explanation is that combining top features with high correlations doesn't necessarily add additional information to the composite; there could be redundancy among digital features. It also suggests opportunities to further improve the performance of the composite digital measure by enriching the feature set with different assessments/tasks and measures.

The superior performance observed in the multivariate analysis, albeit from a small pilot dataset, suggests promises for use of composite digital measures for progression tracking in future studies. Recent modelling efforts have shown that an increased precision made possible by more objective and frequent composite digital measures could lead to smaller and shorter proof-of-concept studies to demonstrate disease-modifying treatment effect (20), which is critical in enabling and accelerating drug development. Open discussions on methodologies to identify the relevant digital features (from the multitude of digital measure possible with DHTs) and construct composite digital measures are critical to enable the adoption of such digital measures, and we present a methodology for this herein.

We see broad applicability of our proposed framework in handling high-dimensional, longitudinal DHT datasets and developing novel digital biomarkers for disease progression. To gain confidence in the use of such biomarkers for decision-making in clinical development, we anticipate that further efforts in technical validation and clinical validation will also be needed to build confidence in the constructed composite digital measures. Additionally, operational feasibility and user acceptance are critical to ensure that the measure can be successfully collected in clinical trials. All these elements will be part of the evidence package to support the fit-for-purpose use of a new digital biomarker and will be important to both clinical implementation and the interpretation of results.

Lastly, although we propose here a machine learning-based approach to develop composite digital biomarkers as indicators of traditional clinical endpoints, it is also valuable to further explore the clinical and biological relevance of the identified features. For example, one could examine the univariate associations between individual digital features and the clinical scores or domain sub-scores (as included in our pipeline and illustrated in Figure 3). Further, the relevance of many symptom features to the underlying disease mechanism have also been reported in the literature. In our illustrative example, several turning features, including peak velocity, duration, and angle, were identified to be valuable for progression tracking; coincidentally, turning has been highlighted in many prior publications as a common challenge in PD (39–42) and is also included in a Phase 2 interventional study as a key digital feature (38).

There are several limitations of our work. First, a major caveat of the results reported from the illustrative example herein is that this analysis only used a small number of participants. Our proposed analysis workflow for digital biomarker development needs to be applied to additional studies with larger N to further demonstrate utility. The identified individual digital features of Parkinson's disease progression and the composite digital measure presented herein is solely for purposes of illustrating the methodology approach. They would need to be validated and verified in an independent dataset in further research before they can be used as digital biomarkers of disease progression and treatment response. Second, the digital features in our study were obtained from sensor-based movement measurements using one DHT system used during supervised, in-clinic tasks. Different or expanded digital features may be available with different DHTs, different task-based assessments, use of passive monitoring approaches, technology evolution, and further algorithm development. It is worth noting that we mainly use this feature set to demonstrate the methodology, and our proposed high-dimensional longitudinal data analysis framework (including feature selection and predictive modelling) is adaptive for different feature sets collected from different sensor technologies. Third, in our illustrative example, we examined potential confounders (i.e., age and sex) in the pooled PD group to identify factors that might significantly impact the progression trending. As a result, age was identified and included into our model as a covariate to be adjusted. It would have been preferred to assess potential confounders in each PD subgroup (i.e., *de novo* PD, mild-to-moderate PD, and advanced PD) rather than the pooled PD group; however, the small sample size and imbalanced datasets within the subgroups posed challenges to doing so.

Notably, the current dataset is longitudinal but only contains in-clinic visit data. One advantage of DHTs is that they may offer the ability to capture data outside of the clinic much more frequently. Other studies, including the Phase 2 Trial of Anti α-Synuclein Antibody in Early Parkinson's Disease (PASADENA) study (10) (daily tasks) and the Personalized Parkinson Project (PPP) study (43) (bi-weekly tasks), have shown utility in capturing remotely collected DHT data with increased measurement frequency. Increased measurement frequency could further enhance the performance of digital measures in quantifying disease progression, as it could address the day-to-day symptom fluctuations and reduce the measurement variability. Such remotely acquired digital features could also be applied to the methodology and framework we've reported here.

In addition, there is emerging research into characterization of the neurodegenerative disease progression directly from raw sensor signals recorded by DHTs (e.g., wearable sensors, environmental sensors, smartphone sensors) using deep neural networks and other black box algorithms (44, 45). Germane to these efforts is an important question about the interpretability of the ensuing

models and results (46, 47). In our work, we identified candidate digital features of disease progression using inherently interpretable linear models. We did not explore deep learning of the raw sensor data directly; such an approach is an interesting future direction of research.

In summary, with the rapid development of DHTs, digital measures are playing an increasingly important role in not only neurodegenerative disease detection, but also longitudinally tracking disease progression over time and detection of therapeutic response. Our proposed ML-based framework for identifying digital features of progression and constructing composite digital measures adds to the existing body of literature on digital measure analysis methodologies and may help accelerate the translation of digital measures to utility for drug development and clinical practice.

Data availability statement

The datasets presented in this article are not readily available because the original data presented in this paper is from the ongoing OxQUIP study and cannot be shared until completion of the whole study and full dissemination of results. This is expected to become possible within 24 months from the end of the study. Qualified researchers will be able to contact the Principal Investigator at the University of Oxford. Requests to access these datasets should be directed to Chrystalina A. Antoniades, chrystalina.antoniades@ndcn.ox.ac.uk.

Ethics statement

The studies involving humans were approved by a research ethics committee and the Health Research Authority (REC16/SW/0262). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

SZ: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal Analysis, Methodology, Software, Visualization. AL: Conceptualization, Formal Analysis, Methodology, Writing – review & editing, Visualization. JS: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. YX: Data curation, Formal Analysis, Writing – review & editing. VS: Conceptualization, Methodology, Writing – review & editing. JF: Funding acquisition, Investigation, Writing – review & editing. CA: Funding acquisition, Investigation, Writing – review & editing. DH: Methodology, Writing – review & editing. MD: Supervision, Writing – review & editing. JR: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Methodology, Supervision. RB: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Methodology, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. J.J.F and C.A.A were supported by the National Institute for Health Research Oxford Biomedical Research Centre. J.J.F has received consulting fees from Abbott and Medtronic, unrelated to this study. J.J.F and C.A.A have received research grant support from UCB Pharma and MSD Laboratories. We thank the participants and their families for their endless support with our research work.

Conflict of interest

SZ, AL, JS, YX, VS, DH, MD, JR and RB were employed by Merck & Co., Inc. JS is a review editor for Statistical Genetics and Methodology in Frontiers in Genetics. JF is an Associate Editor for Neuroprosthetics in Frontiers in Neuroscience. CA is a review editor for Perception Science in Frontiers in Neuroscience and Frontiers in Psychology. JR is a review editor for Soft Matter Physics in Frontiers in Physics. RB is a review editor for Medical Physics and Imaging in Frontiers in Physiology and Frontiers in Physics.

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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2024.1500811/full#supplementary-material

References

1. 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. *Movement Disord.* (2008) 23(15):2129–70. doi: 10.1002/mds.22340

2. Committee for Medicinal Products for Human Use. Qualification Opinion on Dopamine Transporter Imaging as an Enrichment Biomarker for Parkinson's Disease Clinical Trials in Patients with Early Parkinsonian Symptoms. European Medicines Agency:EMA/CHMP SAWP/765041/2017 (2018).

3. Horak FB, Mancini M. Objective features of balance and gait for Parkinson's disease using body-worn sensors. *Mov Disord.* (2013) 28(11):1544–51. doi: 10.1002/mds.25684

4. Heldman DA, Espay AJ, LeWitt PA, Giuffrida JP. Clinician versus machine: reliability and responsiveness of motor endpoints in Parkinson's disease. *Parkinsonism Relat D.* (2014) 20(6):590–5. doi: 10.1016/j.parkreldis.2014.02.022

5. LeBaron V, Hayes J, Gordon K, Alam R, Homdee N, Martinez Y, et al. Leveraging smart health technology to empower patients and family caregivers in managing cancer pain: protocol for a feasibility study. *Jmir Res Protoc.* (2019) 8(12):e16178. doi: 10.2196/16178

6. Evers LJ, Raykov YP, Krijthe JH, de Lima ALS, Badawy R, Claes K, et al. Real-life gait performance as a digital biomarker for motor fluctuations: the Parkinson@home validation study. *J Med Internet Res.* (2020) 22(10):e19068. doi: 10.2196/19068

7. Erb MK, Karlin DR, Ho BK, Thomas KC, Parisi F, Vergara-Diaz GP, et al. Mhealth and wearable technology should replace motor diaries to track motor fluctuations in Parkinson's disease. *NPJ Digitl Med.* (2020) 3(1):6. doi: 10.1038/s41746-019-0214-x

8. 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(1):5. doi: 10.1038/s41746-019-0217-7

9. Burq M, Rainaldi E, Ho KC, Chen C, Bloem BR, Evers LJW, et al. Virtual exam for Parkinson's disease enables frequent and reliable remote measurements of motor function. *NPJ Digit Med.* (2022) 5(1):65. doi: 10.1038/s41746-022-00607-8

10. Pagano G, Boess FG, Taylor KR, Ricci B, Mollenhauer B, Poewe W, et al. A phase II study to evaluate the safety and efficacy of prasinezumab in early Parkinson's disease (PASADENA): rationale, design, and baseline data. *Front Neurol.* (2021) 12:705407. doi: 10.3389/fneur.2021.705407

11. Brzezicki MA, Conway N, Sotirakis C, FitzGerald JJ, Antoniades CA. Antiparkinsonian medication masks motor signal progression in *de novo* patients. *Heliyon.* (2023) 9(11):e16415. doi: 10.1016/j.heliyon.2023.e16415

12. Pagano G, Taylor KI, Anzures-Cabrera J, Marchesi M, Simuni T, Marek K, et al. Trial of prasinezumab in early-stage Parkinson's disease. *New Engl J Med.* (2022) 387(5):421–32. doi: 10.1056/NEJMoa2202867

13. Robin J, Xu M, Detke M, Simpson W. Validation of an objective, SpeechBased object content score for measuring disease progression in AD. *J Prev Alzheimer's Dis.* (2022) 9:S190. doi: 10.14283/jpad.2022.97 (P143).

14. Liu Y, Zhang G, Tarolli CG, Hristov R, Jensen-Roberts S, Waddell EM, et al. Monitoring gait at home with radio waves in Parkinson's disease: a marker of severity, progression, and medication response. *Sci Transl Med.* (2022) 14(663):eadc9669. doi: 10.1126/scitranslmed.adc9669

15. Yang Y, Yuan Y, Zhang G, Wang H, Chen YC, Liu Y, et al. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nat Med.* (2022) 28(10):2207–15. doi: 10.1038/s41591-022-01932-x

16. Kadirvelu B, Gavriel C, Nageshwaran S, Chan JPK, Nethisinghe S, Athanasopoulos S, et al. A wearable motion capture suit and machine learning predict disease progression in Friedreich's ataxia. *Nat Med.* (2023) 29:86–94. doi: 10.1038/s41591-022-02159-6

17. Ricotti V, Kadirvelu B, Selby V, Festenstein R, Mercuri E, Voit T, et al. Wearable full-body motion tracking of activities of daily living predicts disease trajectory in Duchenne muscular dystrophy. *Nat Med.* (2023) 29(1):95–103. doi: 10.1038/s41591-022-02045-1

18. Johnson SA, Karas M, Burke KM, Straczkiewicz M, Scheier ZA, Clark AP, et al. Wearable device and smartphone data quantify ALS progression and may provide novel outcome measures. *NPJ Digit Med.* (2023) 6(1):34. doi: 10.1038/s41746-023-00778-y

19. Giboin LS, Simillion C, Rennig J, Bamdadian A, Kinsella F, McColgan P, et al. A digital motor score for sensitive detection of disease progression in early manifest Huntington's disease (2023). Available online at: https://medically.roche.com/global/en/medical-material.a86d594d-0db2-42d9-9bef-6622e0b574d2.qr.html?cid=slprxx2304nehdchdi2023 (Accessed December 19, 2024).

20. Mori H, Wiklund SJ, Zhang JY. Quantifying the benefits of digital biomarkers and technology-based study endpoints in clinical trials: project moneyball. *Digit Biomark.* (2022) 6(2):36–46. doi: 10.1159/000525255

21. Adams JL, Kangarloo T, Gong Y, Khachadourian V, Tracey B, Volfson D, et al. Using a smartwatch and smartphone to assess early Parkinson's disease in the WATCH-PD study over 12 months. *NPJ Park.'s Dis.* (2024) 10:112. doi: 10.1038/s41531-024-00721-2

22. Czech MD, Badley D, Yang L, Shen J, Crouthamel M, Kangarloo T, et al. Improved measurement of disease progression in people living with early Parkinson's disease using digital health technologies. *Commun. Med.* (2024) 4(1):49. doi: 10.1038/s43856-024-00481-3

23. Perumal SV, Sankar R. Gait and tremor assessment for patients with Parkinson's disease using wearable sensors. *Ict Express.* (2016) 2(4):168–74. doi: 10.1016/j.icte.2016.10.005

24. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS One.* (2016) 11(8):e0157077. doi: 10.1371/journal.pone.0157077

25. Alaskar H, Hussain A. Prediction of Parkinson disease using gait signals. *2018 11th International Conference on Developments in ESystems Engineering (DeSE)* (2018). p. 23–6

26. Gao C, Sun H, Wang T, Tang M, Bohnen NI, Müller ML, et al. Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. *Sci Rep.* (2018) 8(1):7129. doi: 10.1038/s41598-018-24783-4

27. Tsoulos IG, Mitsi G, Stavrakoudis A, Papapetropoulos S. Application of machine learning in a Parkinson's disease digital biomarker dataset using neural network construction (NNC) methodology discriminates patient motor Status. *Frontiers Ict.* (2019) 6:10. doi: 10.3389/fict.2019.00010

28. Dadu A, Satone V, Kaur R, Hashemi SH, Leonard H, Iwaki H, et al. Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *NPJ Park.'s Dis.* (2022) 8(1):172. doi: 10.1038/s41531-022-00439-z

29. Taylor K, Lipsmeier F, Volkova-Volkmar E, Rukina D, Anzures-Cabrera J, Essioux L, et al. Prasinezumab reduced progression of Parkinson's disease motor features measured by roche PD Mobile application v2 sensor features: PASADENA phase II part 1 [abstract]. *Mov Disord.* (2021) 36(suppl 1). https://www.mdsabstracts.org/abstract/prasinezumab-reduced-progression-of-parkinsons-disease-motor-features-measured-by-roche-pd-mobile-application-v2-sensor-features-pasadena-phase-ii-part-1/

30. Sotirakis C, Su Z, Brzezicki MA, Conway N, Tarassenko L, FitzGerald JJ, et al. Identification of motor progression in Parkinson's disease using wearable sensors and machine learning. *NPJ Park.'s Dis.* (2023) 9(1):142. doi: 10.1038/s41531-023-00581-2

31. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* (1986) 73(1):13–22. doi: 10.1093/biomet/73.1.13

32. Wang L, Zhou J, Qu A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics.* (2012) 68(2):353–60. doi: 10.1111/j.1541-0420.2011.01678.x

33. Pereira MF, Buchanan T, Höglinger GU, Bogdanovic M, Tofaris G, Prangnell S, et al. Longitudinal changes of early motor and cognitive symptoms in progressive supranuclear palsy: the OxQUIP study. *BMJ Neurology Open.* (2022) 4:e000214. doi: 10.1136/bmjno-2021-000214

34. Svetnik V, Liaw A, Tong C, Wang T. Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Roli F, Kittler J, Windeatt T, editors. *Multiple Classifier Systems. MCS 2004. Lecture Notes in Computer Science, vol 3077.* Berlin, Heidelberg: Springer (2004). doi: 10.1007/978-3-540-25966-4_33

35. Zampieri C, Salarian A, Carlson-Kuhta P, Aminian K, Nutt JG, Horak FB. The instrumented timed up and go test: potential outcome measure for disease modifying therapies in Parkinson's disease. *J Neurology Neurosurg Psychiatry.* (2010) 81(2):171. doi: 10.1136/jnnp.2009.173740

36. Koop MM, Ozinga SJ, Rosenfeldt AB, Alberts JL. Quantifying turning behavior and gait in Parkinson's disease using mobile technology. *Ibro Reports.* (2018) 5:10–6. doi: 10.1016/j.ibror.2018.06.002

37. Lowry K, Woods T, Malone A, Krajek A, Smiley A, Van Swearingen J. The figure-of-8 walk test used to detect the loss of motor skill in walking among persons with Parkinson's disease. *Physiother Theory Pract.* (2022) 38(4):552–60. doi: 10.1080/09593985.2020.1774948

38. Lipsmeier F, Taylor KI, Postuma RB, Volkova-Volkmar E, Kilchenmann T, Mollenhauer B, et al. Reliability and validity of the Roche PD Mobile application for remote monitoring of early Parkinson's disease. *Sci Rep-uk.* (2022) 12(1):12081. doi: 10.1038/s41598-022-15874-4

39. Stack E, Ashburn A. Dysfunctional turning in Parkinson's disease. *Disabil Rehabil.* (2008) 30(16):1222–9. doi: 10.1080/09638280701829938

40. Hulbert S, Ashburn A, Robert L, Verheyden G. Narrative review of turning deficits in people with Parkinson's disease. *Disabil Rehabil.* (2014) 37(15):1382–9. doi: 10.3109/09638288.2014.961661

41. Mak MK, Patla A, Hui-Chan C. Sudden turn during walking is impaired in people with Parkinson's disease. *Exp Brain Res.* (2008) 190(1):43–51. doi: 10.1007/s00221-008-1446-1

42. Ragothaman A, Miranda-Dominguez O, Brumbach BH, Giritharan A, Fair DA, Nutt JG, et al. Relationship between brain volumes and objective balance and gait measures in Parkinson's disease. *J Park.'s Dis*. (2021) 12(1):283–94. doi: 10.3233/JPD-202403

43. Bloem BR, Marks WJ, de Lima ALS, Kuijf ML, van Laar T, Jacobs BPF, et al. The personalized Parkinson project: examining disease progression through broad biomarkers in early Parkinson's disease. *BMC Neurol*. (2019) 19(1):160. doi: 10.1186/s12883-019-1394-3

44. Atri R, Urban K, Marebwa B, Simuni T, Tanner C, Siderowf A, et al. Deep learning for daily monitoring of Parkinson's disease outside the clinic using wearable sensors. *Sensors*. (2022) 22(18):6831. doi: 10.3390/s22186831

45. Khan P, Fazlul Kader MD, Riazul Islam SM, Rahman AB, Sshahriar Kamal MD, Uddin Toha M, et al. Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances. *IEEE Access*. (2021) 9:37622–55. doi: 10.1109/ACCESS.2021.3062484

46. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: fundamental principles and 10 grand challenges. *Stat Surv*. (2022) 16:1–85. doi: 10.1214/21-SS133

47. US Food and Drug Administration. Good machine learning practice for medical device development: guiding principles (2021). Available online at: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (Accessed December 19, 2024).

frontiers | Frontiers in Digital Health

TYPE Policy and Practice Reviews
PUBLISHED 23 October 2025
DOI 10.3389/fdgth.2025.1415202

OPEN ACCESS

EDITED BY
Toshiyo Tamura,
Waseda University, Japan

REVIEWED BY
Anupama Ginige,
Western Sydney University, Australia
Graham Jones,
Tufts Medical Center, United States

*CORRESPONDENCE
Derek L. Hill
✉ derek.hill@ucl.ac.uk

RECEIVED 10 April 2024
ACCEPTED 25 September 2025
PUBLISHED 23 October 2025

CITATION
Hill DL, Carroll C, Belfiore-Oshan R and Stephenson D (2025) Aligning with regulatory agencies for the use of digital health technologies in drug development: a case study from Parkinson's disease.
Front. Digit. Health 7:1415202.
doi: 10.3389/fdgth.2025.1415202

Aligning with regulatory agencies for the use of digital health technologies in drug development: a case study from Parkinson's disease

Derek L. Hill[1*], Camile Carroll[2,3], Ramona Belfiore-Oshan[4] and Diane Stephenson[4]

[1]Biomedical Engineering, University College London, London, United Kingdom, [2]Translational and Clinical Research Institute, Newcastle University, Newcastle, United Kingdom, [3]National Institute for Health and Care Research (NIHR) Newcastle Biomedical Research Centre (BRC), Newcastle University, Newcastle, United Kingdom, [4]Critical Path Institute, Tucson, AZ, United States

Digital Health Technologies (DHTs) have been under investigation for many years as innovative tools for Parkinson's disease motor symptoms given their inherent high-frequency, sensitive, and objective measurement properties. DHTs used in drug development, can be defined as Drug Development Tools (DDT), though some DHTs may also be categorized as medical devices. The recent rapid increase in use of DHTs in clinical trials has been accompanied by a rapidly evolving regulatory landscape, resulting in a challenging environment for widespread implementation of DHTs in applications that will provide clear impact on pharmaceutical company drug development pipelines. Parkinson's disease represents a disease of escalating burden with high unmet need for therapies that are disease modifying. Early intervention is a key area of focus, yet the heterogeneity of symptoms and lack of biomarkers poses challenges for drug development. Furthermore, the technologies and device platforms, both hardware and software, are rapidly evolving, and the companies developing the underlying devices frequently have objectives and timelines that may not align with those of the pharmaceutical industry. DHTs therefore have a unique set of challenges in terms of devising meaningful measures, standardization of data collected, responding to evolving regulatory expectations, and ensuring alignment across stakeholders. There is a growing need for new models of collaboration to bring together diverse stakeholders required to achieve regulatory endorsement of DHTs for use as DDTs. Collaborations between stakeholders working on DHTs need to be firmly anchored in the regulatory ecosystem as many regulatory challenges in DHTs have parallels in other technologies. Furthermore, there is an especially urgent need to define the pre-competitive space in which DHT data can be shared, data collection standards devised, and novel analysis approaches that are robust to residual variability developed. Critical Path for Parkinson's Consortium's (CPP) Digital Drug Development Tool (3DT) initiative is highlighted as a case example to illustrate how pre-competitive public private partnerships can advance the regulatory maturity of digital health technology measures for use in clinical trials.

KEYWORDS

drug development, digital health technologies, data sharing, regulatory framework, Parkinson's disease, neurological disorders

1 Introduction

Digital Health Technologies (DHTs) used as Drug Development tools (DDTs) represent an important example of a regulated technology to support medical product development. These technologies have the potential to meet pharmaceutical industry needs for high frequency, sensitive, and objective measures of a patient's disease progression, and a patient's response to treatment in real-world settings (1).

DHTs have attracted particular interest in chronic progressive diseases of the nervous system (2). This is due to the heterogenous nature of symptoms, slow insidious onset of symptoms with long duration of presymptomatic underlying disease, and lack of patient centered measures that can be used to define true impact of novel therapies on patient's quality of life.

DHT measures may therefore accelerate the development of new drug and biological therapies in areas of unmet medical need and enable these treatments to be better focused on treating the aspect(s) of disease of most importance to patients.

DHTs, when used to support drug development, sit at the interface between medicine and device regulations. The applicable regulatory landscape is rapidly evolving including across regulatory authorities. Here we make use of terminology from the FDA's recent guidance document on Digital Health Technologies for Remote Data Acquisition (3), and limit our discussion to DHTs that incorporate sensors (e.g., motion sensors). We use the term "DHT measure" to mean the output of a DHT used as a drug development tool, and "DHT Device" for the data collection device e.g., wearable sensor from which the DHT measure is obtained or derived.

In this paper, we describe the work of Critical Path Institute's (C-Path) Critical Path for Parkinson's (CPP) Digital Drug Development Tools (3DT) initiative to collect evidence that DHTs can reliably and accurately measure PD progression at early stages in drug naïve patients over one year duration, in order to advance the regulatory maturity of DHTs for assessing patients with Parkinson's disease (PD). CPP is a public private partnership focused on the development of drug development tools targeting early stages of the disease. The key milestones for DHTs being used as DDTs are (a) letter of support and (b) qualification. A letter of support is issued by the medicines regulator to describes the regulator's thoughts on the potential value of a DDT and encourages further evaluation. A DDT qualification is a public regulatory opinion that encourages the use of a qualified DDT for a specific context of use to expedite drug development and review of regulatory applications.

The regulatory landscape for DHTs has been evolving significantly since CPP was launched in 2015 (4): there has been a rapid increase in the response of regulators to the needs of DHTs and their use in drug development. Regulatory agencies have published several guidance and discussion documents focused on DHTs and with some DHT measures reaching a high level of maturity with certain regulators. This regulatory framework enables DHTs to be used on a protocol-specific basis, or to be qualified for more general application in a context of use. Many DHT measures are generated using machine learning (ML) and artificial intelligence (AI), which means they may be impacted by AI-specific regulations being proposed in several jurisdictions, including the European Union AI Act (5).

Table 1 shows the timeline for advances in the regulatory landscape over the past several years both in U.S. and Europe, with key regulatory publications from the U.S. Food and Drug Administration (FDA) and European Medicines Agency (EMA) highlighted. Although primarily focused on regulation of medicinal products, we include cybersecurity guidance documents focused on medical devices generally relevant to all DHTs, whether medical devices or not. Notably, the FDA's March 2024 "AI & Medical Products" guidance specifically describes how medicines and device regulators are working together in this rapidly evolving area.

2 Unique challenges of DHTs

Regulators have made much progress in provision of guidance for DHTs in drug development, though the impact of DHTs in clinical trials has so far been limited; for example, no drug has yet been approved by the FDA based on a DHT derived primary endpoint (6) and the EMA has recently described regulatory experience with DHTs in the context of registrational studies as minimal (7). Issues relate to the rapid rate of innovation in digital technologies, the types of companies in the ecosystem, and the intersection between regulations related to clinical trials, medical devices, and data protection/privacy.

2.1 Rapid rate of innovation

The rapid rate of innovation in the technologies incorporated in DHTs (e.g., sensors, ML algorithms, connected devices) means that the product lifecycle of a DHT is often a small number of years. A DHT may rely on consumer computing platforms such as smartphones. The lifetime of DHT devices, and sometimes even digital companies, is short compared to the timescale of drug development. It is therefore hard for DHTs to "travel with a molecule" from phase I to approval, which might be a period of more than 10 years. Even if a particular hardware remains stable, the installed software might periodically upgrade in ways that make the data non-comparable.

2.2 Standardization and harmonization

A consequence of the rapid rate of innovation in the hardware, software and measurements from DHTs is the need to obtain comparable data across time and studies. The diversity in technologies available, the speed of innovation, including software upgrades and new versions of hardware, and the proprietary nature of some algorithms means that obtaining comparable data is a considerable challenge.

One state-of-the-art approach in this area has been described by the Mobilise-D consortium (8), in which multiple types of motion sensors have been compared against a gold standard in

TABLE 1 Recent key regulatory guidance and frameworks relevant to DHTs. The majority are published by medicines regulators though cybersecurity guidance documents published by medical device regulators are also included.

Date	Regulator	Title	Comment	Link
June 2018	FDA	Patient-Focused Drug Development: Collecting Comprehensive and Representative Input	Guidance document	https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-collecting-comprehensive-and-representative-input
June 2020	EMA – Human Medicines Division	Questions and Answers: Qualification of Digital Technology-Based Methodologies to Support Approval of Medicinal Products	Document to support qualification of DHT methodologies	https://www.ema.europa.eu/en/documents/other/questions-and-answers-qualification-digital-technology-based-methodologies-support-approval-medicinal-products_en.pdf
July 2020	EMA – Medical Device Coordination Group	MDCG 2019–16 Rev.1 Guidance on Cybersecurity for medical devices	Guidance Document	https://health.ec.europa.eu/document/download/b23b362f-8a56-434c-922a-5b3ca4d0a7a1_en
February 2022	FDA	Patient-Focused Drug Development: Methods to Identifying What Is Important to Patients	Guidance document	https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-methods-identify-what-important-patients
June 2022	FDA	Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments	Guidance document	https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-selecting-developing-or-modifying-fit-purpose-clinical-outcome
April 2023	FDA	Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into	Guidance document	https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-incorporating-clinical-outcome-assessments-endpoints-regulatory
May 2023	FDA - CDER	Artificial Intelligence for Drug Development	Informational	https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/artificial-intelligence-drug-development
May 2023	FDA - CDER	Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products	Discussion Paper/Request for Feedback	https://www.fda.gov/media/167973/download
March 2023	FDA	Framework for the Use of Digital Health Technologies in Drug and Biological Product Development	Framework; PDUFA VII	https://www.fda.gov/media/166396/download?attachment
March 2023	EMA – GCP IWG	Guideline on computerised systems and electronic data in clinical trials	Guidance Document	https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-computerised-systems-and-electronic-data-clinical-trials_en.pdf
Sept 2023	FDA	Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions	Guidance Document	https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-system-considerations-and-content-premarket-submissions
Dec. 2023	FDA	Digital Health Technologies for Remote Data Acquisition in Clinical Investigations	Guidance Document	https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations
Jan 2025	FDA	Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products	Draft Guidance	https://www.fda.gov/media/184830/download
June 2025	FDA	Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions	Guidance document	https://www.fda.gov/media/119933/download

a laboratory setting. This highlighted considerable remaining challenges in standardizing DHT data even from accelerometers, which are arguably the most mature of DHT sensor technology. The authors suggest guidelines to assist standardization efforts for future studies.

Parallels have previously been drawn between DHTs and imaging. Putting in place suitable standardization has been important in the development of neuroimaging in clinical trials (9, 10) and is a focus of the FDA guidance on imaging endpoints in clinical trials (11). It is important to note that, while there are parallels with imaging, DHTs are used for remote data acquisition (e.g., in the home) and there is considerable additional variability compared to that of the in-clinic controlled environment applicable to imaging. This puts additional requirements on the standardization of DHTs that allow for bridging in-clinic with at-home measurements. Standardization of a particular DHT measure, therefore, should consider implications of hardware, software, and measurement environment. The experience of standardizing imaging endpoints encourages the standardization to be done in the context of a specific measurement such as hippocampal volume (12) or Positron Emission Tomography standardized uptake value (PET SUV) (10), and for measurements obtained from diverse scanners (sometimes with contrast or tracers) and algorithms to be compared in terms of effect size in a relevant comparison e.g., separating diseased from normal or progressing from non-progressing subjects (13, 14). Once a measurement is clearly defined, the standardization task is easier to specify. The lack of consensus on specific DHT measures has been

a barrier to progress in this measurement-driven standardization. Because some DHT devices can generate multiple possible DHT-derived measures (for example the output from a wrist-worn accelerometer could be used to calculate measures of gait, tremor or sleep), the appropriate standardization and algorithm validation should be measure rather than device specific.

2.3 Business models and data protection and privacy

Technology companies, whether focused on digital health or consumer tech, frequently have business models that involve monetizing data (15). Sophisticated consumer hardware and software used by individuals is provided at low cost (and for software, often free) in exchange for the user agreeing to transfer their data to the tech company and give ownership, or at least wide-ranging rights to use it for commercial purposes. The huge volumes of data thus acquired by the tech companies can be used to improve the product, but also can be sold freely, so an individual's data may be used by unknown third parties for purposes that were neither pre-defined nor specifically consented to by the user. These data-centric business models are potentially incompatible with the desire of pharmaceutical companies, healthcare providers, and regulators to ensure that patient data is carefully controlled and only used for pre-specified purposes with informed consent.

2.4 Intersection between different regulatory systems

A further challenge relates to DHTs operating at the interface between different regulatory frameworks. Many DHT devices (e.g., smartphones and smartwatches with fitness apps and activity trackers) are designed for consumer use and have limited regulatory oversight. A sub-set of DHT devices are either medical devices or contain software components that are "software as a medical device". Use of any of these DHT devices in clinical trials adds new regulatory requirements around validation of computer systems that come from Good Clinical Practice (GCP) (16) (21CFR11 in USA, Annex 11 of the Clinical Trial Regulation in Europe). The EMA has made clear in recent publications that GCP regulations around validation and audit trail apply to mass market wearables and mobile phones (17). Some digital health companies struggle to put in place systems that are compliant with these requirements and do not see a business case for achieving compliance, given the small size of the clinical trial market for most of these companies.

The need for different models of data use, and the requirements of validation and audit trail, mean that commercial collaborations between the pharmaceutical and tech sectors can be challenging.

This further emphasizes that for DHTs to have a significant impact on the development of new treatments, new models of collaboration are needed. There is also a need to acknowledge that the price point of the technologies used in clinical trials is likely to be significantly higher than the prices that end-users are used to for consumer digital technologies.

3 The need for new models of collaboration to develop DHTs

In recent years, there has been significant optimism that "digital" technologies could rapidly impact drug development, and as a result, relevant industry and public organizations are investing in DHTs across various therapeutic areas. There has been an associated rapid increase in the number of clinical studies incorporating DHTs (2), particularly in chronic progressive disorders of the nervous system where the failure rate is high and there is a lack of sensitive, clinically meaningful DDTs. The application of DHTs to disorders of the nervous system is growing at a rapid rate with Parkinson's being most prominent of all (Evidence from https://www.ClinicalTrials.gov on the growth of Digital Health Technologies in neurology trials (2).

It is increasingly clear that while DHTs have great potential to positively impact drug development, the timescale of their development has not proved to be rapid in comparison to other technologies such as imaging, and at the date of writing, we have not yet seen any new drugs approved based on a DHT measurement. One DHT measure that has achieved the regulatory milestone of being qualified as a primary endpoint in Duchenne muscular dystrophy (DMD) by the EMA is the Stride Velocity 95th centile (SV95C) (18). This effort took more than a decade (19) to complete, which is not indicative of the minimum (or maximum) time required but illustrates the challenges of navigating the regulatory environment for DHTs. While most recent DMD studies have included SV95C as a secondary outcome (NCT05524883, NCT05096221, NCT06138639, NCT05982119, NCT04906460), the use of this measure has been explored for other neuromuscular diseases including Spinal Muscular Atrophy, Facioscapulohumeral muscular dystrophy, and Limb Girdle muscular dystrophy. However, it is still unclear how the learnings from the DMD qualification will be applied or whether they are fully translatable to those other diseases (20).

Many pharmaceutical companies and research institutions have been independently working on developing DHT measures, which has resulted in an explosion of proposed approaches to measuring concepts of interest such as gait (21). It is becoming increasingly clear that the challenges are too big to overcome as individual companies and organizations alone, necessitating a collaborative and harmonized approach. Increasingly, pharmaceutical companies are looking for a clear impact on their drug development programs and adapting their investment in DHTs accordingly. A consortium-based approach is therefore desirable and aligns with regulatory agency recommendations for public-private partnerships to increase their efficiency in advancing DHTs (22, 23). Some industry-led consortia have sought to develop high-impact DHT measures that are disease-agnostic or are cross-disease digital endpoints in areas such as

fatigue, sleep (24), and mobility (21). Regulators, however, have consistently communicated that, just as for other (non-digital) technologies, data should be submitted for a single disease and context of use (COU).

It is therefore increasingly important that, for reasons of cost effectiveness and rate of progress, development of DHTs is undertaken collaboratively rather than in isolation, and anchored within organizations that have wide-ranging experience in development of non-DHT DDTs. Some of the DHT challenges identified above could be addressed by means of collaborative data analysis platforms such as federated learning.

4 The evolving DHT regulatory landscape

While DHTs have been used in clinical research for decades (25), there has been significant increase in use over the last 5 years particularly post-COVID-19 pandemic, and a rapid evolution in the regulatory landscape for DHTs as DDTs. In particular, there are recent regulatory publications specific to DHTs (3, 26) and those that can apply to DHTs including those on patient-focused drug development, use of AI in devices (27), drug development (28), and validation of computer systems (17).

Industry has proposed the use of DHTs for several applications in drug development that span a variety of different intended uses to enhance decision making in clinical trials, not only as digital endpoints (29). DHTs have potential to be used for advancing novel candidate therapies at all stages of drug development including patient subgroup characterization, optimizing trial design, patient identification and recruitment, risk assessment and adverse event prevention, remote interventions to enable decentralized clinical trials, externally controlled trials, and label indication expansion.

Up until 5 years ago, it was common to refer to all DHT measures as "digital biomarkers". However, the DHT measures can be used for multiple purposes to support drug development, and as such, the use of DHTs might meet either the definition of a biomarker or of a clinical outcome assessment (30):

- **Digital Biomarker**: "a characteristic or set of characteristics, collected from digital health technologies, that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions." (31)
- **Clinical outcome assessment (COA)**: an assessment of how someone feels, functions, or survives (32).

For some DHT measures, this distinction remains a matter of debate (33). For example, it is possible to argue that change in a measure of gait due to progression or treatment of PD is both "an indicator of a pathogenic process or biological response" and that it is an "assessment of how someone feels or functions". This distinction has practical applications. For a "biomarker", the sensitivity of the measure to the pathogenic process or biological response is the priority, with the goal of achieving a larger effect size and hence needing fewer participants and/or less time for a clinical trial for a new medicine, in which demonstration of drug efficacy is the objective. For COA, however, clinical meaningfulness is the priority, and a sensitive measure that is not meaningful to the participant or their physician would be considered inappropriate in a trial in which the objective is demonstration of clinical effectiveness. This has implications for the types of data needed to advance the regulatory maturity of DHTs. The next section discusses the regulatory focus on patient-focused drug development, which is of great relevance to the use of DHTs for COAs.

4.1 Patient-focused drug development

Medicines regulators have an increasing focus on ensuring that data collected during clinical trials of new medicines takes account of the patients' voice. The FDA's recent series of guidance documents on patient-focused drug development (34–37) refer to DHTs in various places, and it is clear that regulators will treat many DHT measures as a type of Clinical Outcome Assessment (COA). The implication for the use of DHTs in clinical trials is that regulators want to see evidence that the DHT measure is relevant to a meaningful aspect of health for the patient. For example, accelerometers have become ubiquitous for tracking activity in smartphones and smartwatches. There are established ways of calculating "activity metrics" from this acceleration data, e.g., step count, cadence and amount of vigorous activity, and many novel motion-sensor-derived measures can be developed using machine learning and artificial intelligence. The focus on meaningfulness of DHT measures means that it is necessary to show that the DHT measure can be linked to a concept of interest relevant to the condition, and a meaningful aspect of patient health. This approach is being followed by consortia working in some disease areas e.g., nocturnal scratch (38). This linkage between DHT measure and meaningful aspects of health needs to be shown for each clinical condition, and regulatory agencies have similar expectations as to data required for drug development tools such as biomarkers and COAs (e.g., both observational and clinical trial data to support a defined COU).

Regulators are using the term "fit for purpose" to describe when a DHT measure is ready for use in a clinical investigation, and they make clear that a DHT measure has to be validated for a single COU,: it is considered fit-for-purpose when "the level of validation associated with a medical product development tool is sufficient to support its context of use" (30).

Whether a DHT is fit for purpose is determined by the strength of the evidence in support of interpreting the DHT measure as reflecting the concept of interest within the COU. Fit-for-purpose in the regulatory context means the same thing as valid within modern validity theory, e.g., validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (39).

4.2 FDA digital health technology guidance/framework

In 2021, the FDA published a draft guidance, "Digital Health Technologies for Remote Data Acquisition in Clinical Investigations" (40), and subsequently published a framework document that seeks to explain how DHTs fit into FDA's thinking (26); a final version of the DHT guidance was published in December 2023 (3). Key implications of this guidance are that the initial step in choosing an appropriate DHT is to "consider the clinical event or characteristic of the disease or condition of interest that is to be measured, identify appropriate technical and performance specifications of a DHT, and consider the proposed trial population". In practice, very often innovation in DHTs has started with available DHT devices (e.g., wrist-worn accelerometers) and sought to derive from this DHT device a DHT measure that meets a drug development need. This guidance further emphasizes the need to clearly define a rationale for the selection of a particular DHT for a context of use, the need for appropriate verification, validation, usability assessment, and the consideration of risks, including confounds (they give the example of false positive detection of tremor in PD from a person traveling in a car on a bumpy road). In the framework published, the FDA acknowledged that it needs to adapt internally to be able to properly consider DHTs and provide sponsors with consistent feedback between review divisions.

4.3 Machine learning and AI in drug development

Many DHT measures are calculated using machine learning (ML) or artificial intelligence (AI). Developers and users of DHTs therefore need to take account of the evolving regulatory landscape for AI. This is an area of rapid evolution in regulatory thinking and a potentially significant divergence between jurisdictions. The FDA has recently published a discussion paper "Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products" (28), which is relevant to DHTs. Of particular relevance is the need to manage risk that arises from use of ML/AI models, which the regulators argue can be distinct from risk in traditional rules-based software. These risks include data quality risks, bias risks (e.g., selection bias, confounding variables), and data security and privacy risks (41).

4.4 Recent DHT regulatory milestones

As of August 2025 there are a total of two letters of support and two full qualification opinions from the EMA on the use of DHTs as drug development tools as digital endpoints. The FDA manages a public website (41) showing it has accepted multiple digital endpoints into the COA qualification program for a range of conditions including DMD, Multiple Sclerosis (MS), chronic heart failure, sarcopenia and atopic dermatitis. By reviewing the Agency feedback provided in each case example there are common issues to be addressed even though the specific indication may be different (42). Sharing of such knowledge and learnings promises to catalyze progress and avoid redundancies and inefficiencies.

5 Critical path institute's 3DT initiative

C-Path is a not-for-profit organization that has nearly two decades of experience leading public-private partnerships spanning multiple diseases to advance regulatory maturity of drug development tools (Table 2) across several neurological disorders including Alzheimer's disease (AD), PD, and DMD. C-Path-led consortia have achieved regulatory milestones from full qualification opinions to Letters of Support and Fit for Purpose FDA and EMA endorsements (43).

The 3DT initiative in Parkinson's disease was launched in 2018 under the auspices of the established global consortium, CPP, as a data-driven collaborative path to share knowledge and resources. The vision of 3DT is to advance the regulatory maturity of DHTs as drug development tools for decision-making in PD trials targeting early Parkinson's disease.

CPP's 3DT initiative has provided a data-driven framework for multiple sponsors who have agreed to collaborate on optimizing the use of DHTs in PD drug development. The 3DT consortium involves sharing of patient-level digital device data (including raw data) with members. The 3DT consortium has maintained regular interaction with medicines regulators, including a Critical Path Innovation Meeting (CPIM) held with the FDA and an Innovation Task Force (ITF) meeting with EMA, both in 2019. Regular additional interactions include with FDA staff members regularly attending monthly consortium meetings, thereby providing an ongoing regulatory dialogue. These interactions with regulators have highlighted several challenges facing the field, including the need for strategies for establishing meaningful clinical endpoints, controlling sources of variability, and evaluating DHT performance in normative as well as diseased cohorts.

A key focus of CPP 3DT is the observational study WATCH-PD (Wearable Assessment in the Clinic and at Home in PD) (NCT03681015) which is focused on an early *de novo* PD target population. This study evaluates the ability of research-grade wearable sensors, a smartwatch and a smartphone to assess key features of PD, using a platform that maps directly onto the MDS-UPDRS. WATCH-PD aims to determine the specific disease features these digital tools can detect, whether the

TABLE 2 Critical path institute (C-path) regulatory milestones to date.

Regulator	Letters of support		Qualifications	
	Total	% led by C-Path	Total	% led by C-Path
FDA	25	44%	16	50%
EMA	49	20%	30	30%

FDA, US Food and Drug Administration; EMA, European Medicines Agency, as of March 2024.

measures differed between individuals with early PD and age-matched controls, and how well the digital measures correlated with traditional ones (44, 45). The CPIM and ITF meetings in 2019 provided regulatory feedback that was used to refine the Watch-PD protocol, adding a normal control arm, and including more rigorous qualitative evaluation of the meaningfulness of the DHT measures to study participants, illustrating the value of early interaction with regulators. CPP recognizes that WATCH-PD is a single study that is noninterventional and has limitations.

5.1 3DT progress to date

3DT has brought together a group of leading industry partners, academic key opinion leaders, patient advocacy organizations, and people living with PD from around the world.

The key components and milestones in the phases of 3DT are shown in Table 3.

6 Discussions and conclusions

There is an evolving regulatory landscape for Digital Health Technologies as drug development tools, with multiple stakeholders independently approaching regulatory agencies for endorsement. Experience of many parallel initiatives approaching regulatory agencies to date suggests that navigating the regulatory path to enable DHTs to have a significant impact on drug development and defining success in addressing drug development needs remain challenging. The experiences of the 3DT consortium highlight the value of collaborative approaches involving pharma industry and academic experts, leveraging Critical Path Institute's experience of advancing the regulatory maturity of a diverse range of drug development tools, from Patient Reported Outcomes (PROs) to imaging biomarkers (46). Tackling challenges collectively by advancing data-driven solutions and sharing costs and risks, as well as embracing open science, can avoid duplication of effort and therefore improve the efficiency with which we advance the regulatory acceptance of DHTs and their use in clinical trials. While DHTs make use of different technologies from those used in other DDTs, C-Path's experience in other types of DDTs, and its existing infrastructure for legal, data, and regulatory engagement has proved valuable in enabling the 3DT consortium to progress. Specific regulatory feedback on the Watch-PD case study itself (such as the need to incorporate a control group, and to add a qualitative element to the study to assess the symptoms of most importance to patients) has informed multiple sponsors as to which considerations are essential across device platforms, both in other PD applications and in different disease areas.

The experiences to date make clear that, while digital technologies have many distinct characteristics, the use of DHT measures as drug development tools needs to fit into the same framework as other DDT technologies. It is therefore essential to precisely define:

- The concept of interest (COI): a clinical event or characteristic of the disease or condition of interest that is to be measured, as either a COA or biomarker.
- The application of the DHT In terms of how it will be applied for drug development decision making (COU). The way the DHT measures the COI will impact the drug development process.
- The rationale for the use of a particular DHT measure relevant to that COI including why it meets the required technical and performance specifications.
- How the selected DHT measure is meaningful.
- The evidence that demonstrates the DHT measure is sufficiently well validated for the COU ("fit for purpose").

For much work to date on DHT- measures as DDTs, it is hard to precisely define all these elements. A diversity of stakeholders is key to success and spans technology experts, clinicians, industry, academic experts, nonprofit organizations, people with lived experience, and regulators themselves. New approaches and new models of collaboration are needed to advance the field as

TABLE 3 Key components of C-path's CPP 3DT.

Regulatory alignment	Data strategy	Patient focused approach	Legal framework
Formal engagement with FDA (CPIM) and EMA (ITF and qualification advice)	C-Path platform for curation and sharing of DHT data, including raw sensor data, within consortia	Included PD-affected individual in WATCH-PD study design	Informed consent for WATCH-PD included data sharing with C-Path
Informal engagement with FDA and EMA regulators at consortium meetings and workshops including joint with EFPIR	Sharing of unprocessed in-clinic and at-home WATCH-PD data while study on-going.	Shared patient-centric trial recommendations using DHT	Data sharing agreements in place with consortium members and C-Path advisors.
Role of C-Path consortia highlighted at 4 workshops hosted by regulators	Sharing of DHT data from pharma sponsored studies.	Data from qualification study shared with patients.	HIPAA and GDPR compliance
Feedback from regulators impacted Watch-PD protocol and analysis plans including addition of control arm	Anonymised Data available to individual sponsors for research and development use only (not commercialization)		
Co-authored abstracts and manuscripts	Analysis design takes account of regulatory feedback.		

EMA, European Medicines Agency; FDA, Food and Drug Administration; EFPIA, European Federation of Pharmaceutical Industries and Associations; ITF, Innovative Task Force; CPIM, critical path innovation meeting; HIPAA, health insurance portability and accountability Act; GDPR, general data protection regulation.

TABLE 4 Nine recommended next steps to take the field forward.

#	Recommendation
1	Define a pre-competitive space in which pharmaceutical companies, device companies, academic experts and people with lived experience can collaborate on specific COIs and COUs.
2	Ensure alignment of incentives for all stakeholders, taking account of differing business models and the need to devise tools that can be deployed in settings with low network bandwidth, limited digital literacy, and in low and middle-income countries.
3	Build on this alignment within the pre-competitive space to enable meaningful sharing of DHT data for defined regulatory purposes, taking into account ethical and pragmatic considerations.
4	Establish good practice for demonstrating meaningfulness of DHT-derived measures.
5	Establish good practice for demonstrating equivalence between different hardware/software for a given DHT measure.
6	Devise standardization approaches in data acquisition, how devices are used in studies, data handling, and data analysis for defined DHT measurements for a COI and catalyze the implementation of these in future studies.
7	Develop collaborative data analytics platforms that are able to handle the large data volumes collected and are designed to be robust to residual variation in data collection given the rapidly evolving and heterogenous nature of DHT hardware and embedded software.
8	Provide a clearer roadmap for demonstrating "fit for purpose" DHTs by focusing on some exemplar measures. Align across parallel consortia to advance multiple data sources synergistically.
9	Define pathways to improve usability to reduce patient and site burden, especially in diverse and global clinical trial populations.

efficiently as possible to be able to attend to the time-sensitive needs of patients. Such collaborative approaches should learn lessons from other types of DDTs (e.g., imaging) to address challenges of standardization and collaborative implementation of analysis methods to enable convergence rather than divergence of proposed DHT measurements. Given the challenges of integrating and harmonizing legacy data collected across distinct device platforms, it is recommended that precompetitive collaborations focus on sharing risks, costs, and prospective study design and collection to optimize DHT studies for the future. We propose nine crucial next steps to advance the field, as shown in Table 4. While these recommendations are based on experience with this Parkison's disease case study, they are more generally applicable for DHTs used as DDTs in this regulatory environment.

Author contributions

DH: Writing – original draft, Conceptualization, Visualization, Validation, Methodology, Writing – review & editing, Data curation. CC: Writing – original draft, Methodology, Conceptualization, Writing – review & editing. RB-O: Writing – review & editing, Validation, Visualization. DS: Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. DH receives funding from UKRI grant numbers 10113073 and 10110484 to support his research in regulatory science. Critical Path Institute is supported by the Food and Drug Administration (FDA) of the Department of Health and Human Services (HHS) and is 54% funded by the FDA/HHS, totaling $19,436,549, and 46% funded by non-government source(s), totalling $16,373,368.

Acknowledgments

The authors acknowledge Critical Path Institute staff Kimberly Ward Barowicz for her contributions to figures and Laura Carrillo, Erin Lowry and Shasta Jorgensen for their review of the article. We recognize our academic collaborators and PIs of the WATCH-PD study, Jamie Adams, Jennifer Mammen, and Ray Dorsey. The industry firms Biogen and Takeda are commended for their willingness to share data and collaborate with Critical Path Institute. We want to thank the members of the 3DT initiative for generously funding and collaborating to support the 3DT project and a talented group of scientific advisors who have supported this effort. CPP acknowledges Scientific advisors Camille Carroll, Jesse Cedarbaum, Tanya Simuni, George Roussos, Anat Mirelman, Luc Evers, Marjan Meinders, Bas Bloem and Lisa Shipley. Members include Biogen, Takeda, AbbVie, Novartis, Merck, UCB, GSK, Roche, The Michael J. Fox Foundation and Parkinson's UK. We recognize the contributions of Critical Path Institute staff in supporting this initiative. The authors specifically acknowledge and appreciate the unique participation of FDA and EMA representatives who have shaped this project. Critical Path Institute is supported by the Food and Drug Administration (FDA) of the Department of Health and Human Services (HHS) and is 54% funded by the FDA/HHS, totaling $19,436,549, and 46% funded by non-government source(s), totalling $16,373,368. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS or the U.S. Government.

Conflict of interest

The 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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

1. Warmerdam E, Hausdorff JM, Atrsaei A, Zhou Y, Mirelman A, Aminian K, et al. Long-term unsupervised mobility assessment in movement disorders. *Lancet Neurol*. (2020) 19(5):462–70. doi: 10.1016/S1474-4422(19)30397-7

2. Masanneck L, Gieseler P, Gordon WJ, Meuth SG, Stern AD. Evidence from ClinicalTrials.gov on the growth of digital health technologies in neurology trials. *NPJ Digit Med*. (2023) 6(1):23. doi: 10.1038/s41746-023-00767-1

3. Digital Health Technologies for Remote Data Acquisition in Clinical Investigations Guidance for Industry, Investigators, and Other Stakeholders. (2023). Available online at: https://www.fda.gov/media/155022/download (Accessed April 1, 2024)

4. Stephenson D, Alexander R, Aggarwal V, Badawy R, Bain L, Bhatnagar R, et al. Precompetitive data sharing as a catalyst to address unmet needs in Parkinson's disease. *J Park Dis*. (2015) 5(3):581. doi: 10.3233/JPD-150570

5. Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future. (2021). Available online at: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (Accessed March 29, 2024)

6. Bakker JP, Izmailova ES, Clement A, Hoffmann S, Leptak C, Menetski JP, et al. Regulatory pathways for qualification and acceptance of digital health technology-derived clinical trial endpoints: considerations for sponsors. *Clin Pharmacol Ther*. (2025) 117(1):56–72. doi: 10.1002/cpt.3398

7. Jadoenathmisier KD, Gardarsdottir H, Mol PGM, Pasmooij AMG. Insights from the European medicines agency on digital health technology derived endpoints. *Drug Discov Today*. (2025) 30(6):104388. doi: 10.1016/j.drudis.2025.104388

8. Palmerini L, Reggi L, Bonci T, Del Din S, Micó-Amigo ME, Salis F, et al. Mobility recorded by wearable devices and gold standards: the mobilise-D procedure for data standardization. *Sci Data*. (2023) 10(1):38. doi: 10.1038/s41597-023-01930-9

9. Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging*. (2008) 27(4):685–91. doi: 10.1002/jmri.21049

10. Klunk WE, Koeppe RA, Price JC, Benzinger TL, Devous MD Sr, Jagust WJ, et al. The centiloid project: standardizing quantitative amyloid plaque estimation by PET. *Alzheimers Dement*. (2015) 11(1):1–15.e4. doi: 10.1016/j.jalz.2014.07.003

11. Research C for DE and. Clinical trial imaging endpoint process standards guidance for industry. (2020). Available online at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-imaging-endpoint-process-standards-guidance-industry (Accessed March 28, 2024)

12. Yu P, Sun J, Wolz R, Stephenson D, Brewer J, Fox NC, et al. Operationalizing hippocampal volume as an enrichment biomarker for amnestic mild cognitive impairment trials: effect of algorithm, test-retest variability, and cut point on trial cost, duration, and sample size. *Neurobiol Aging*. (2014) 35(4):808–18. doi: 10.1016/j.neurobiolaging.2013.09.039

13. Risacher SL, Saykin AJ, Wes JD, Shen L, Firpi HA, McDonald BC. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr Alzheimer Res*. (2009) 6(4):347–61. doi: 10.2174/156720509788929273

14. Ortner M, Drost R, Hedderich D, Goldhardt O, Müller-Sarnowski F, Diehl-Schmid J, et al. Correction to: amyloid PET, FDG-PET or MRI?—the power of different imaging biomarkers to detect progression of early Alzheimer's disease. *BMC Neurol*. (2020) 20(1):80. doi: 10.1186/s12883-020-01649-9

15. Ofulue J, Benyoucef M. Data monetization: insights from a technology-enabled literature review and research agenda. *Manag Rev Q*. (2022) 74:521–65. doi: 10.1007/s11301-022-00309-1

16. ICH harmonised guideline integrated addendum to ICH E6(R1): Guideline for Good Clinical Practice ICH E6(R2) ICH Consensus Guideline. ICHGCP. Available online at: https://ichgcp.net/home (Accessed March 28, 2024)

17. Notice to sponsors on validation and qualification of computerised systems used in clinical trials. clinical trials.

18. Final Qualification Opinion for Stride velocity 95 centile as primary endpoint in studies in ambulatory Duchenne Muscular Dystrophy studies.

19. Servais L, Camino E, Clement A, McDonald CM, Lukawy J, Lowes LP, et al. First regulatory qualification of a novel digital endpoint in duchenne muscular dystrophy: a multi-stakeholder perspective on the impact for patients and for drug development in neuromuscular diseases. *Digit Biomark*. (2021) 5(2):183–90. doi: 10.1159/000517411

20. Poleur M, Markati T, Servais L. The use of digital outcome measures in clinical trials in rare neurological diseases: a systematic literature review. *Orphanet J Rare Dis*. (2023) 18(1):224. doi: 10.1186/s13023-023-02813-3

21. Micó-Amigo ME, Bonci T, Paraschiv-Ionescu A, Ullrich M, Kirk C, Soltani A, et al. Assessing real-world gait with digital technology? Validation, insights and recommendations from the mobilise-D consortium. *J Neuroeng Rehabil*. (2023) 20(1):78. doi: 10.1186/s12984-023-01198-5

22. Cerreta F, Ritzhaupt A, Metcalfe T, Askin S, Duarte J, Berntgen M, et al. Digital technologies for medicines: shaping a framework for success. *Nat Rev Drug Discov*. (2020) 19(9):573–4. doi: 10.1038/d41573-020-00080-6

23. Sacks L, Kunkoski E. Digital health technology to measure drug efficacy in clinical trials for Parkinson's disease: a regulatory perspective. *J Parkinsons Dis*. (2021) 11(s1):S111–5. doi: 10.3233/JPD-202416

24. Nobbs D, Piwko W, Bull C, et al. Regulatory qualification of a cross-disease digital measure: benefits and challenges from the perspective of IMI consortium IDEA-FAST. *Digit Biomark*. (2023) 7(1):132–8. doi: 10.1159/000533189

25. Meijer GAL, Westerterp KR, Verhoeven FMH, Koper HBM, ten Hoor F. Methods to assess physical activity with special reference to motion sensors and accelerometers. *IEEE Trans Biomed Eng*. (1991) 38(3):221–9. doi: 10.1109/10.133202

26. Framework for the Use of DHTs in Drug and Biological Product Development. (2023). Available online at: https://www.fda.gov/media/166396/download (Accessed October 8, 2024)

27. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. Silver Spring: Food and Drug Administration (2024). Available online at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (Accessed November 7, 2024)

28. Using Artificial Intelligence & Machine Learning in the Development of Drug and Biological Products. (2023). Available online at: https://www.fda.gov/media/167973/download (Accessed March 28, 2024).

29. Biogen-Regulatory-Policy-White-Paper_Transforming-the-Regulatory-Landscape-for-Digital-Health-Technologies-in-Drug-Developmen.pdf. Available online at: https://www.biogen.com/content/dam/corporate/international/global/en-US/docs/pdfs/Biogen-Regulatory-Policy-White-Paper_Transforming-the-Regulatory-Landscape-for-Digital-Health-Technologies-in-Drug-Developmen.pdf (Accessed March 28, 2024)

30. FDA-NIH Biomarker Working Group. *BEST (Biomarkers, EndpointS, and Other Tools) Resource*. Silver Spring: Food and Drug Administration (US) (2016). Available online at: http://www.ncbi.nlm.nih.gov/books/NBK326791/ (Accessed March 28, 2024)

31. Vasudevan S, Saha A, Tarver ME, Patel B. Digital biomarkers: convergence of digital health technologies and biomarkers. *NPJ Digit Med*. (2022) 5:36. doi: 10.1038/s41746-022-00583-z

32. FDA. Clinical Outcome Assessment (COA): Frequently Asked Questions. Available online at: https://www.fda.gov/about-fda/clinical-outcome-assessment-coa-frequently-asked-questions (Accessed October 8, 2024)

33. Izmailova ES, Demanuele C, McCarthy M. Digital health technology derived measures: biomarkers or clinical outcome assessments? *Clin Transl Sci*. (2023) 16(7):1113–20. doi: 10.1111/cts.13529

34. Patient-Focused Drug Development: Collecting Comprehensive and Representative Input Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. (2020). Available online at: https://www.fda.gov/media/139088/download (Accessed December 6, 2023)

35. Patient-Focused Drug Development: Methods to Identify What Is Important to Patients Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. (2022). Available online at: https://www.fda.gov/media/131230/download (Accessed December 6, 2023)

36. Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for Purpose Clinical Outcome Assessments Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders DRAFT GUIDANCE. (2022). Available online at: https://www.fda.gov/media/159500/download (Accessed December 6, 2023)

37. Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders DRAFT GUIDANCE. (2023). Available online at: https://www.fda.gov/media/166830/download (Accessed March 28, 2024).

38. Cesnakova L, Meadows K, Avey S, Barrett J, Calimlim B, Chatterjee M, et al. A patient-centred conceptual model of nocturnal scratch and its impact in atopic dermatitis: a mixed-methods study supporting the development of novel digital measurements. *Skin Health Dis*. (2023) 3(5):e262. doi: 10.1002/ski2.262

39. Research C for DE and. Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments. (2022). Available online at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-selecting-developing-or-modifying-fit-purpose-clinical-outcome (Accessed March 28, 2024)

40. Digital Health Technologies for Remote Data Acquisition in Clinical Investigations. (2021). Available online at: https://www.fda.gov/media/155022/download (Accessed November 10, 2023)

41. New AAMI Consensus Report: Guidance on Risk Management for AI, ML. News. (2022). Available online at: https://array.aami.org/do/10.2345/article.7e472902-8552-45cc-b869-59efe8f7eeb6/ (Accessed March 28, 2024)

42. Stephenson D, Alexander R, Aggarwal V, Badawy R, Bain L, Bhatnagar R, et al. Precompetitive consensus building to facilitate the use of digital health technologies to support Parkinson disease drug development through regulatory science. *Digit Biomark*. (2020) 4(Suppl 1):28–49. doi: 10.1159/000512500

43. Stephenson D, Belfiore-Oshan R, Karten Y, Keavney J, Kwok DK, Martinez T, et al. Transforming drug development for neurological disorders: proceedings from a multidisease area workshop. *Neurotherapeutics*. (2023) 20(6):1682–91. doi: 10.1007/s13311-023-01440-x

44. Mammen JR, Lerner A, Al-Rubayie R, Kostrzebski M, Stephenson D, Xiao Y, et al. Longitudinal qualitative assessment of meaningful symptoms and relevance of WATCH-PD digital measures for people with early Parkinson's. *J Neurol*. (2025) 272(2):114. doi: 10.1007/s00415-024-12789-0

45. Adams JL, Kangarloo T, Gong Y, Khachadourian V, Tracey B, Volfson D, et al. Using a smartwatch and smartphone to assess early Parkinson's disease in the WATCH-PD study over 12 months. *NPJ Parkinsons Dis*. (2024) 10(1):112. doi: 10.1038/s41531-024-00721-2

46. Karpen SR, White JK, Mullin AP, O'Doherty I, Hudson LD, Romero K, et al. Effective data sharing as a conduit for advancing medical product development. *Ther Innov Regul Sci*. (2021) 55(3):591. doi: 10.1007/s43441-020-00255-8

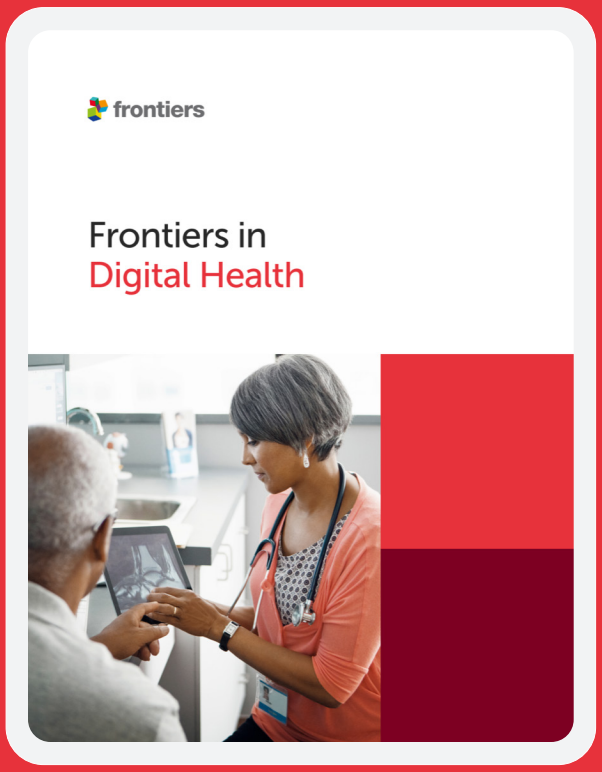
Frontiers in
Digital Health

Explores digital innovation to transform modern healthcare

A multidisciplinary journal that focuses on how we can transform healthcare with innovative digital tools. It provides a forum for an era of health service marked by increased prediction and prevention.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact