ARTICLE OPEN

Identifying and characterising sources of variability in digital outcome measures in Parkinson's disease

George Roussos[1]✉, Teresa Ruiz Herrero[2], Derek L. Hill[3], Ariel V. Dowling[4], Martijn L. T. M. Müller[5], Luc J. W. Evers[6], Jackson Burton[7], Adrian Derungs[8], Katherine Fisher[7], Krishna Praneeth Kilambi[7], Nitin Mehrotra[9], Roopal Bhatnagar[5], Sakshi Sardar[5], Diane Stephenson[5], Jamie L. Adams[10], E. Ray Dorsey[10] and Josh Cosman[11]

Smartphones and wearables are widely recognised as the foundation for novel Digital Health Technologies (DHTs) for the clinical assessment of Parkinson's disease. Yet, only limited progress has been made towards their regulatory acceptability as effective drug development tools. A key barrier in achieving this goal relates to the influence of a wide range of sources of variability (SoVs) introduced by measurement processes incorporating DHTs, on their ability to detect relevant changes to PD. This paper introduces a conceptual framework to assist clinical research teams investigating a specific Concept of Interest within a particular Context of Use, to identify, characterise, and when possible, mitigate the influence of SoVs. We illustrate how this conceptual framework can be applied in practice through specific examples, including two data-driven case studies.

npj Digital Medicine (2022)5:93 ; https://doi.org/10.1038/s41746-022-00643-4

INTRODUCTION

Intensified by the implications of the COVID-19 era[1], Digital Health Technologies (DHT) are widely recognised as a promising complementary element in the clinical assessment of Parkinson's disease (PD). A key enabler is the wider availability of smartphones and wearables which offer the opportunity to enable monitoring of disease progression in daily life[2–4]. More frequent assessments can provide better insight into episodic disease features such as motor fluctuations, freezing of gait, and falls, while avoiding observation bias[5]. Yet, to operationalize DHTs as drug development tools, they must meet the key challenge of regulatory acceptability, so that digital outcome measures can be established as evidence for medical product development.

Yet, despite progress made toward regulatory maturity of DHTs, their use in clinical research is not yet fully accepted[6]. Common challenges in the adoption of DHTs include small study samples, samples that do not reflect accurately the characteristics of the target population, lack of a normative data set, feature selection bias, failure to replicate results due to differences in sensor placement and calibration, and lack of transparency in the use of analytical techniques[7]. When employed at home, DHTs enable higher-frequency data collection compared to traditional clinical assessments. However, this setting can also introduce significantly greater variability between subjects, for example due to differences in apartment size, and within subjects, for example due to differences in room temperature and the presence of family members. Furthermore, studies incorporating machine learning (ML) and artificial intelligence (AI) based approaches in particular, are at high risk of providing overly optimistic results due to feature selection bias when a large number of post hoc candidate features are considered in a relatively limited sample[8]. This is especially relevant when cross validation methods are used to assess performance on a single modestly-sized dataset.

In this context, a key consideration is how to identify, characterise, and when possible, mitigate the influence of key sources of variability (SoVs) introduced by the measurement process and to understand their influence against changes to symptom severity and disease progression. This challenge is further intensified by the heterogeneous nature of PD expression leading to high intra- and inter-study variability.[9] For example, to address a specific hypothesis, selection of study subjects is often biased (e.g., early disease only) and therefore typical variability associated with disease heterogeneity is reduced within a specific study. Ideally, to address this issue multiple data sources would be needed. However, the availability of data sets using DHTs is limited and analyses on limited data can artificially increase the degree of explained variability, leading to bias in insights and predictions. Variability introduced by the measurement process, such as differences in the placement of a wearable, lack of control of the home environment, device software upgrades in the course of a study, or the accuracy of the specific model of sensor used, must be set into the context of normal variability in the subject and how this is impacted by PD.

This paper provides a conceptual framework to assist clinical research teams to identify, characterise and mitigate the influence of key SoVs introduced by the measurement process and contrast their effect against changes due to PD severity and progression. We illustrate how this conceptual framework can be applied in practice through multiple examples including two case studies developed using pilot data contributed by the co-authors.

RESULTS

The primary focus in the design of a clinical investigation is the clinical event or measurable characteristic of PD that is to be assessed and the proposed trial population[10]. For example, the clinical research team would typically identify appropriate

[1]Birkbeck College, University of London, London, UK. [2]Bill and Melinda Gates Foundation, Seattle, WA, USA. [3]Panoramic Digital Health, Grenoble, France. [4]Takeda, Deerfield, IL, USA. [5]Critical Path Institute, Tucson, AZ, USA. [6]Radboud University Medical Center and Radboud University, Nijmegen, The Netherlands. [7]Biogen, Cambridge, MA, USA. [8]Roche, Basel, Switzerland. [9]Alnylam Pharmaceuticals, Cambridge, MA, USA. [10]University of Rochester, Rochester, NY, USA. [11]Abbvie, North Chicago, IL, USA. ✉email: g.roussos@bbk.ac.uk

outcome assessments, preferably a Performance Outcome (PerfO) when DHTs are considered, or digital biomarkers that are meaningful in the specific Context of Use. In this regard, Taylor, et al.[10] outlined the importance of distinguishing between data- and patient-centric approaches: While either approach could influence the assessment of motor experiences in PD due to a variety of SoVs, appropriate mitigation strategies such as test-retest studies are recommended. Next, alternative DHTs should be assessed in terms of design and operation and their suitability considering the education, language, age and technical aptitude of the population targeted. The goal is to establish that the particular device choice is fit-for-purpose for the specific clinical investigation including its physical characteristics; to validate its outputs including data format and accuracy; and to validate the selected digital outcome measure and the method of its calculation. Last but not least, the clinical research team must provide objective evidence that the selected technology and associated measurement process accurately assesses the clinical event or characteristic in the proposed participant population. To this end, investigation of SoVs should be considered a core ingredient in developing comprehensive and convincing evidence of validation, ideally through the quantitative assessment of their influence against performance changes due to Parkinson's.

Identifying and characterising SoVs in the DHT measurement processes

The design of the clinical protocol and of the measurement process may affect SoVs differentially: for example, free-living vs. clinical setting assessments or the choice between active or passive tests. Each alternative may introduce different trade-offs that need to be considered separately. While variability is likely to be greater in a free-living environment, this setting may be preferable to provide greater insight into activities of daily living, therefore better represent the patient's individual capacity, and be less biased as a comparison against a clinical scale. Indeed, a key distinction among DHTs for the assessment of PD is between: (i) active assessments, in which the subject is prompted to perform a particular set of movements, activities, or tasks at a particular time and for a specific duration; and, (ii) passive assessments, where data is sampled continuously by a recording device worn on the body without prompting or other types of direct interaction with the subject. The most common approach in conducting active assessments today involves the use of a smartphone app such as mPower, HopkinsPD, OPDC, Roche, and cloudUPDRS[3,4,11]. These apps typically guide the subject through a series of tasks that are often associated with specific sub-items of Part III motor assessments of the MDS-UPDRS, a rating scale often used clinically to assess for severity of PD features[12] (for examples of typical movements during active testing cf. video at http://www.updrs.net/help/). In any of these studies, the app uses one or more of the smartphone sensors such as accelerometer, gyroscope, microphone, magnetometer, and touch screen, to record measurements associated with the specific task. Apps typically also collect contextual information such as the time of medication intake, self-assessments of well-being, or answers to clinical questionnaires such as the PDQ-8, and may incorporate cognitive assessment tasks such as the Stroop test[13]. In contrast, passive monitoring is used to assess patients based on activities of daily living. Consequently, passive monitoring approaches induce less patient burden compared to active monitoring tasks and support continuous monitoring over days. Moreover, continuous passive monitoring can be used to assess response fluctuations of dopaminergic medication as well as the detection of episodic features, e.g., freezing of gait and falls. Finally, passive monitoring approaches typically fix the placement and orientation of the wearable device, and thus multiple devices are required to assess left and right and upper and lower body movements.

Precept 1: Establish SoVs in active vs passive measurement processes. Measurement processes for active and passive measurements introduce different SoVs. Active tests require the subject to actively engage with a device following a specific schedule. Measurements are influenced by clinical protocol-dependent variations in the number, frequency, and the exact timing of the active tasks performed by the study subjects. Due to the prescribed nature of these measurements, missing data may also occur, which is less likely in a passive measurement process.

In contrast, due to the lack of environmental context in passive testing, it is often challenging to accurately identify the specific task or activity undertaken by the subject during data recording. For example, a type of movement such as riding a bicycle, may not always be adequately recognized. Because it is not always possible to establish ground truth through observation, the practical alternative is often to infer context using machine learning techniques[14]. A common approach is to employ pre-trained models to classify sensor data into activities such as sitting, walking, cooking and so forth. However, such computational methodologies are still in relatively early stages of development, especially at population level, and can accurately account only for a small proportion of all daily activity. Furthermore, manual annotation of activities is still required for validation of algorithm performance. AI and ML algorithms trained to detect the types of activity of clinical interest may then perform poorly when passively collected data contains many types of activities that were not in the original training and validation data. The largest to date published longitudinal study of daily activity achieved less than 30% accuracy across subjects with the best individual accuracy of less than 65%[15]. The key characteristics of active and passive approaches are summarized in Table 1.

Table 1. Comparison between active vs. passive digital assessments.

Active assessments	Passive assessments
Proactive interaction with associated patient burden	Relatively unobtrusive operation with low patient burden
Specific duration of observation	Continuous measurement
Relatively small volume of data	Relatively large volume of data
Known context of data collection constrained to specific movements	Unknown context of data collection affected by unknown external factors
Can be combined with clinical assessments	Predominately unsupervised operation in a non-clinical setting
Effort-intensive to conduct longitudinally	Longitudinal observation by default
Episodic assessment of specific tasks	Real-life functioning of subjects
SoVs can be more easily recognized and examined systematically. Controlling of SoVs is feasible (see also precept 2).	SoVs are more difficult to identify and typically are more difficult to replicate. Controlling of SoVs is less feasible.

Table 2. Mapping SoVs across different measurement process phases.

Data acquisition	Device/sensor configuration
	Assessment tasks and duration
	Sensor positioning and orientation
	Environment
	Schedule of assessment
	Precision and frequency
	Meta-data: device specification, data acquisition setup, file naming, hardware, and software versioning
Data management	Source data file transmission
	Data receipt notification
	Data quality control (missing data, malfunctioning device or sensor, erroneous sampling, erroneous transmission, corrupted storage, timing errors)
	Adverse events assessment
	Notification of data quality concerns and troubleshooting
Data analysis	Signal processing method used for feature extraction
	Signal processing architecture: edge, cloud, or hierarchical/hybrid
	Documentation of algorithms and implementation

The detail of these phases is device and application-specific, for example in some applications, significant data analysis is done on the wearable device itself.

Precept 2: Identify SoVs associated with acquisition, management, and analysis within the measurement process. The measurement process for both active and passive approaches can be separated into three distinct stages and key SoVs relevant to each stage can be identified (cf. Table 2). Detailed descriptions of each factor included in the three distinct phases, namely data acquisition, management, and analysis, are included in a companion paper derived by the work of the Critical Path for Parkinson's Consortium 3DT Working Group[16] on metadata standards and reported in ref. [17].

Precept 3: Characterise low-, medium- and high-impact SoVs. The third element of the conceptual framework characterises SoVs as low, medium, and high impact relative to the risk they present in terms of their potential to cause harm on the ability of digital outcomes to measure clinically relevant aspects of PD if they are not dealt with appropriately. Low-impact SoVs are those that are well-understood and mitigation strategies are readily available, often already incorporated in devices or as a standard feature of data processing software. Medium impact SoVs are well understood and effective means for their control and mitigation are widely available and in common use, for example, through the application of appropriate algorithms or user experience design approaches. Compared to low-impact SoVs, they require more attention, and their mitigation should be specifically addressed in study design but appropriate mitigation measures but do not require extensive further investigation. Finally, high impact SoVs are those that present a significant risk to influence the performance of the digital outcome measure of interest, their characteristics are not adequately documented and quantified, thus mitigation approaches are not readily available or require further validation. Note that the concept of impact in this context incorporates the risk to reduce the fidelity of the measure as well as the maturity and robustness of mitigation methods. However, it excludes the degree of complexity of the mitigation technique applied; for example, low-impact SoVs may still require the implementation of advanced computational methods. Moreover, as discussed later in this paper, note that the precise risk of harm by a particular SoV is only possible to fully quantify within a specific Context of Use.

Low-impact SoVs

Following the above categorisation, thermal effects introducing variability in accelerometer measurements would be classified as a low-impact SoV due to the fact that their effect is well-understood[18] and, indeed, the vast majority of good quality commercial devices incorporate a temperature sensor which is used to adjust the data output accordingly. Gravity is also considered a low-impact SoV for acceleration measurements when the sensor orientation can change. In this case, the effect of gravitation on magnitude estimation can be removed by the application of a standard high-pass filter on the 3-axis signal. When movement directionality along each axis is required, an algorithm such as an L_1-trend filter can be used[19].

A further example of low-impact SoV is the audio capture quality of current smartphones: Grillo et al.[20] tested a variety of devices and found negligible variability in the calculation of common acoustic voice measures using a commercial software tool including many of those widely used in PD[21]. However, they discovered considerable overall differences when alternative algorithms were used to calculate the same measure, suggesting that software artifacts present a higher impact SoV. In this case, algorithm implementation would be considered a medium-impact SoV as it was still possible to mitigate software variability by adjusting for the observed trend across calculated measures, which followed similar patterns.

Medium impact SoVs

Any location of sensor placement, for example at the wrist, foot, ankle, lower back, and chest, offers a distinct trade-off between comfort and variability. For example, if the measurement process requires the estimation of a walking parameter such as speed and stride length, a cumbersome foot-mounted sensor will produce the highest accuracy measurement with the least amount of variability; a lower back or chest-mounted strap would result in high to moderate accuracy and variability; and, a widely available wrist sensor would produce the least accurate and most variable information[22,23]. Overall, variability increases as the distance between sensor and body location of interest increases, which implies that mitigation strategies should aim to minimise separation between the two locations. For example, using a foot rather than wrist sensor when a subject walks while holding a phone, will clearly offer significantly greater accuracy in gait parameter estimation. Nevertheless, practical sensor placement may be influenced by accessibility, subject comfort, and even cultural norms. When the preferred location is not available, careful algorithm selection can help reduce the influence of this SoV[24].

Arguably, next to location the second most influential SoV is the orientation of the sensor on the body. Accelerometers in particular are extremely sensitive to changes in orientation: For example, a typical accelerometer with a range of +/− 8 g and a 10-bit analog-to-digital converter (ADC) will have a resolution of approximately 1.4 degrees. A 10-degree body position change will produce a bias of 0.06 g change in acceleration. Large orientation variations can be expected in practice, sensors are not always precisely placed by the subject, or sensors can erroneously be placed upside down, backward, or at an odd angle, resulting in a large constant bias. While a constant bias can be estimated and removed by a high-pass filter set at a very low frequency, for example, 0.25 Hz, non-constant bias is much more challenging to remove. Non-constant bias due to frequent orientation changes is especially likely to occur when a sensor is attached over clothing

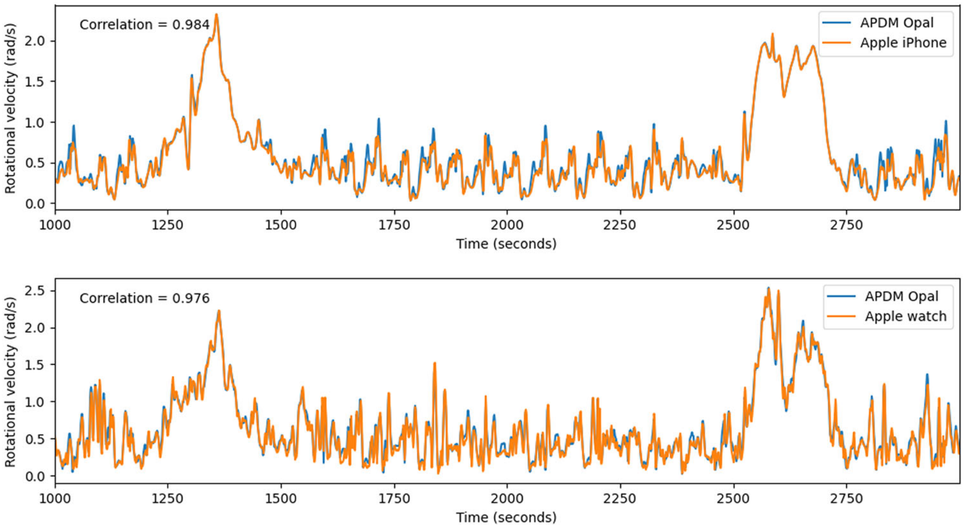

Fig. 1 Angular Velocity Comparison. Top: Comparison of angular velocity calculations using APDM Opal (blue) and iPhone (orange) samples with both devices placed at lumbar region. Bottom: Comparison of angular velocity calculated using Opal (blue) and Apple Watch (orange) samples with both devices placed on the same wrist of the subject.

on a body part with high mobility such as the wrist, or a large muscle group such as the quadriceps femoris, or when a sensor is loosely affixed to the body. This can result in significant localized movement and rotation of the sensor relative to the body during data collection resulting in significant fluctuations in the signal which can significantly lower the signal fidelity. Remediation of this SoV is to ensure that the measurement process provides specific guidance such as all sensors be fixed tightly on the body underneath articles of clothing to minimize relative movement, especially when placed on a hyper-mobile body part, such as the wrist.

Further mitigation of orientation SoVs is possible through the use of orientation-invariant algorithms[25]. A common approach is to first estimate the true sensor orientation on the body, and subsequently calculate the rotational offset between the actual and the “ideal” sensor orientation using standard mathematical transformations[26] and subsequently to employ orientation-invariant correction algorithms. Alternatively, selecting orientation-agnostic features where possible, such as those derived in the frequency domain, would effectively eliminate variability from orientation. Further, it is conceivable to investigate the influence of sensor placement and orientation on sensor data, sensor data features, and digital biomarkers using a novel biomechanical simulations method introduced by Derungs and Amft[27].

High-impact SoVs

While low- and medium-impact SoVs have established mitigation strategies, high-impact SoVs require careful consideration and may require auxiliary exploratory studies to investigate and quantify their influence and hence may require considerable additional effort for the development of appropriate mitigation measures.

Data-driven investigation of SoV impact: case studies

Although it is often possible to use published literature to assess the impact of SoVs, certain settings require the clinical research team to explore specific SoVs within a particular Context of Use and with reference to specific outcome measures. In this Section, we present two case studies that follow a data-driven approach to investigate the potential impact of particular choices in the measurement process. The work presented below is not intended as a comprehensive investigation of the specific SoVs considered, but rather, as a way to illustrate a practical approach to assess their impact at the pilot stage of clinical research. Our analysis is focused on practical ways to identify relevant SoVs of concern before committing to a clinical study protocol design. When specific SoVs are identified as potentially high-impact a full follow-up investigation would be required for example by modelling their impact in terms of erroneous diagnosis.

Case study 1: Device type, number of sensors, and sampling rate

Using data collected during exploratory in-clinic piloting of WATCH-PD (cf. Methods section below and Adams et al.[28], we investigate variability across device types including a popular consumer platform, and differences due to their placement, sampling rate and sampling locations. To compare consumer- (Apple Watch and iPhone) against research-grade (APMD Opal) devices, we analysed data recorded during a one-minute-walk task, where both devices were simultaneously employed (APMD Opal sensors were placed under the Apple device). Figure 1 shows angular velocity calculations obtained from gyroscope data from several subjects, comparing Opal against Apple Watch and iPhone. Opal data were recorded at 128 Hz, down-sampled and time-shifted to align with Apple Watch measurements (bottom) and separately with iPhone (top). Figure 1 suggests that both consumer-grade devices reproduce high, low, and intermediate frequencies at comparable quality to the Opal reference (with correlation of 0.984 and 0.976 correspondingly).

Further, we compared gait features obtained from Opal measurements using the Mobility Lab software provided by APDM (cf. https://apdm.com/mobility/) now part of Clario, against iPhone data processed using software developed in-house (by co-author TRH). The latter, employs the El-Gohary et al.[29] algorithm to identify gait bouts after turns, and subsequently extract gait features using GaitPy[30] following the approach suggested by ref. [31]. In-house developed software (also by TRH) was used to compute rotational velocity at the wrist during arm swings per gait cycle. Figure 2 suggests very strong agreement between the two approaches across all features (with correlation of cadence, gait arm range and turns exceeding 0.9). Figure 3 further suggests that both approaches result into comparable levels of variation in all features. However, gait speed and stride length appear to produce significant (but consistent) differences in absolute terms. This is caused by the

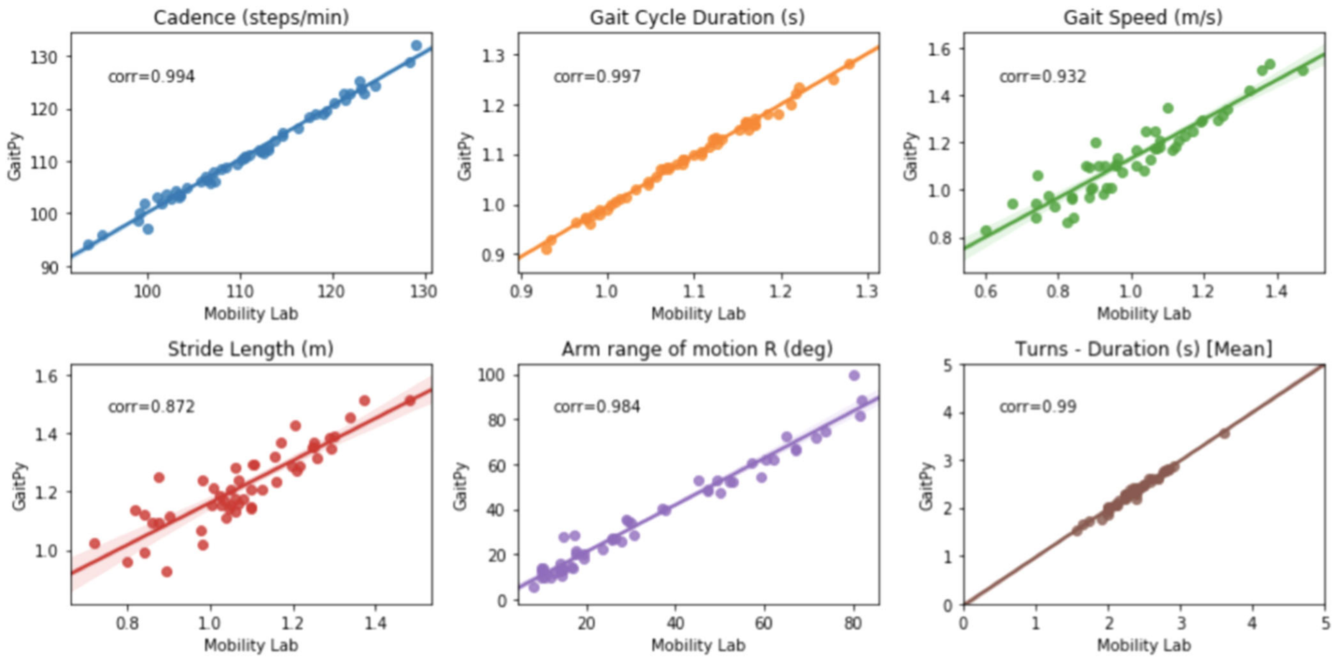

Fig. 2 Gait Features Comparison. Correlation between gait features calculated on the same measurements performed using Opal and using GaitPy and Mobility Lab correspondingly.

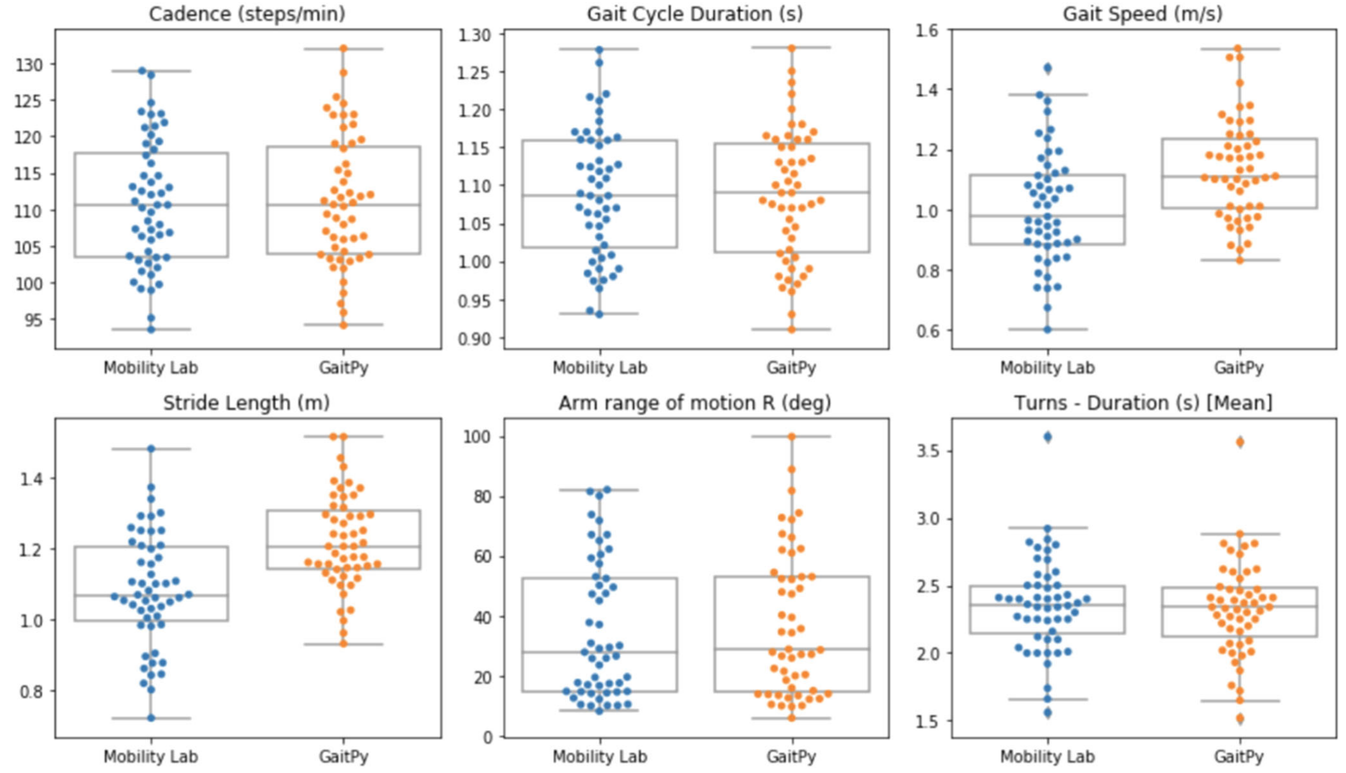

Fig. 3 Gait Feature Variability. Variability of selected gait features calculated on the same measurements performed using Opal and using GaitPy and Mobility Lab correspondingly. The solid line represents the median value; the box limits show the interquartile range (IQR) from the first (Q1) to third (Q3) quartiles; the whiskers extend to the furthest data point within Q1–1.5*IQR (bottom) and Q3 + 1.5*IQR (top).

use of per-subject height and leg-length measurements obtained at enrolment in Mobility Lab calculations, while in GaitPy a fixed height-to-leg-length factor is employed across all subjects. The latter clearly limits the accuracy of the pendular model employed by in the calculation of these features.

Finally, in Fig. 4 gait features estimated using Opal measurements are compared against measurements from consumer-devices using the software developed in-house (signals were aligned as described above). While there is still strong agreement overall, there are also noticeable differences. One cause for this

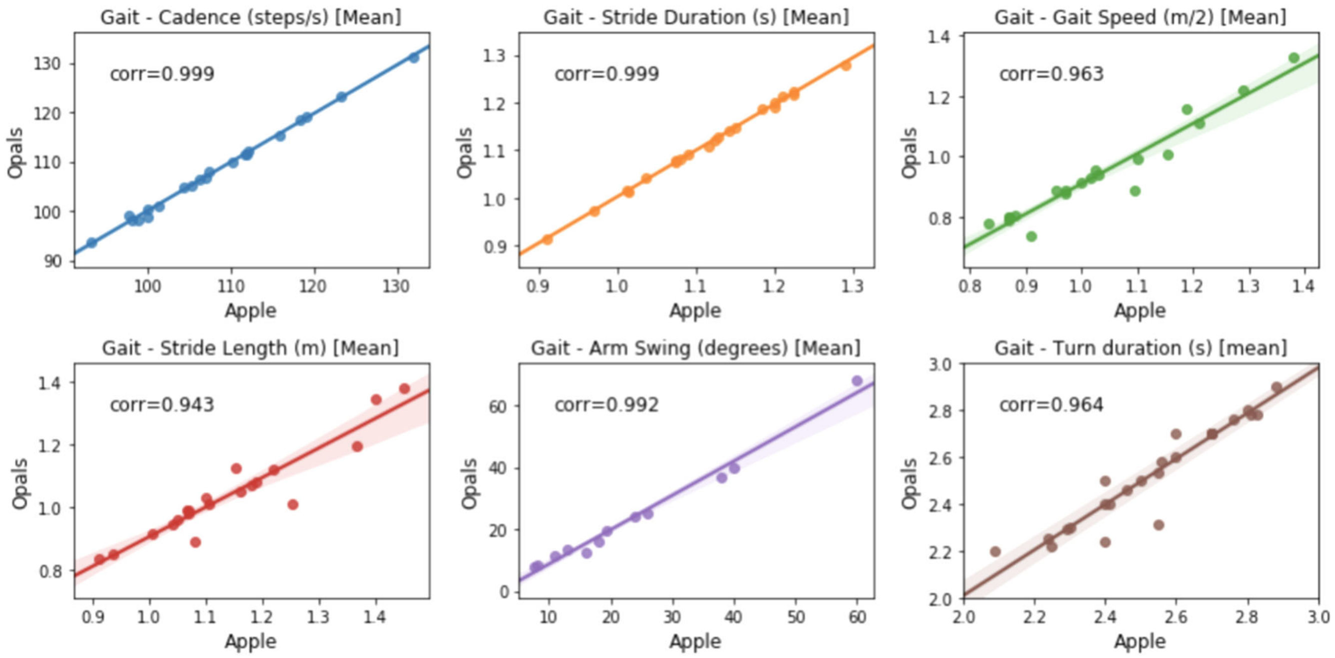

Fig. 4 Comparison against Consumer-grade Devices. Comparing features estimated using Opal versus consumer-grade devices using the software developed in-house. All features use iPhone measurements except arm swing that employs data sampled using the Apple Watch.

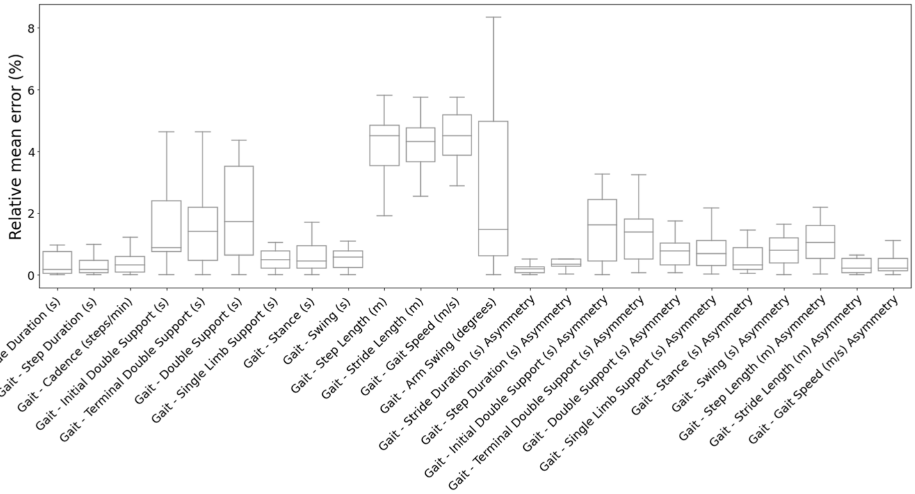

Fig. 5 Variability of Gait Features. Distribution of the relative mean error across 50, 100 and 128 Hz sampling rates per calculated feature. The solid line represents the median value; the box limits show the interquartile range (IQR) from the first (Q1) to third (Q3) quartiles; the whiskers extend to the furthest data point within Q1–1.5*IQR (bottom) and Q3 + 1.5*IQR (top).

mismatch is likely to be due to the small angular misalignment introduced by the specific placement of the devices on top of each other as described above.

To explore sampling frequency as a SoV, Opal measurements were down-sampled to obtain data at 50 and 100 Hz. Figure 5 demonstrates the limited impact of lower sampling rates on feature estimation in terms of error. Features involving double support and asymmetry are most affected because they are more sensitive to error propagation caused by small inaccuracies in the calculation of underlying metrics. This interpretation is supported by the findings of ref. [32], which conducted an extensive evaluation of seven different IMUs: Accelerometer and gyroscope data from each device were processed using the same algorithm and compared against ground truth obtained using OptoGait (cf. http://optogait.com). Similar to our analysis, temporal parameters demonstrated less variability to spatial parameters for which more complex calculations are needed for, example, double integration and an error-state Kalman filter, and are thus sensitive to even small measurement inaccuracies. Zhou et al.[32] traced the latter to device issues such as insufficient ADC range or inadequate sensor calibration. Overall, our investigation suggests that features less sensitive to low-frequency sampling can be identified using the above observations as appropriate for the specific Concept of Interest and Context of Use.

Case study 2: Environmental Factors

Variability due to environmental factors, that is, factors relating to the setting within which the measurement process is performed, rather than the process per se, can also affect outcome measures. For example, Perraudin et al.[33] identified the height of the chair used to perform sit-to-stand transition time tests as a key environmental SoV in this context. Using data provided by the DOMVar project obtained from an actigraphy bracelet incorporating gyroscope and accelerometer (cf. Methods Section below), Fig. 6 illustrates the effect of three chairs of different heights (39.5 cm, 51.5 cm, and 59.5 cm) on average across 12 transitions for each of three subjects. Note the significantly higher variability when data are aggregated across chair heights. Hence, passive monitoring at a patient's home, where typically different chair types would be present, is likely to result into less consistent outcome measure estimation.

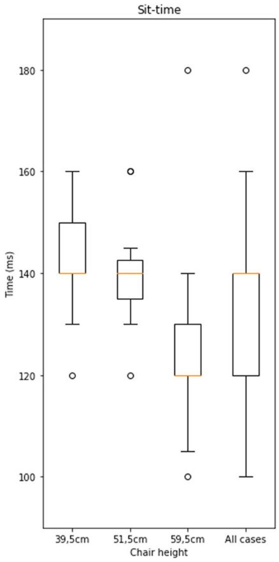

Fig. 6 Variability due to Chair Selection. Chair height as a SoV for sit-to-stand transition time tests. Variability is considerably larger when considered across chairs. The error bars of boxplot are generated by matplotlib using the matplotlib.pyplot.boxplot function with default parameters. The boxplot extends from the first to the third quartile of the data with a line at the median. The whiskers extend from the box by 1.5 times the inter-quartile range.

Further, walking speed can be strongly influenced by room size and the arrangement of furniture within in. Figure 7 illustrates stride time variability for two healthy subjects. Data is collected passively using the actigraphy bracelet in four different settings, namely: (i) large empty room, (ii) large room containing furniture obstacles, (iii) small empty room, and (iv) small room containing furniture obstacles.

Both examples above suggest that passive monitoring, in particular, is especially sensitive to environmental factors. When the passive monitoring is preferable for clinical reasons, averaging the relatively larger number of measurements can reduce variability when no systematic changes in the SoVs are expected, for example, when the layout of the patient's home changes to accommodate further restrictions in their movements their symptoms progress.

DISCUSSION

In this paper, we introduced a conceptual framework for the identification and characterisation of SoVs related to the use of DHTs in clinical trials for PD. We distinguish SoVs related to experimental design and choice of technology against variability introduced by the subject, either inherently or due to the disease. This framework aims to provide practical guidance on how to investigate, assess, and where possible, mitigate their influence on the measurement process targeting a particular Concept of Interest in a specific Context of Use.

To this end, the choices between active or passive monitoring and the duration of the study are especially influential. In our experience, investigators often incorporate elements of both active and passive assessments despite the lack of due justification. Active approaches are often sufficient to provide conclusive evidence and achieve higher specificity of the derived outcomes measures. For example, an active approach to quantify movement quality would be less likely to be affected by environmental factors (Case Study #2). However, passive monitoring would be preferable when the relevant Concept of Interest is associated with the subject's overall patterns of movement, such as general long-term activity levels or the quantification of relatively rare events such as falls and freezing. Indeed, in the case of falls and freezing, active assessment would likely be ineffective despite its

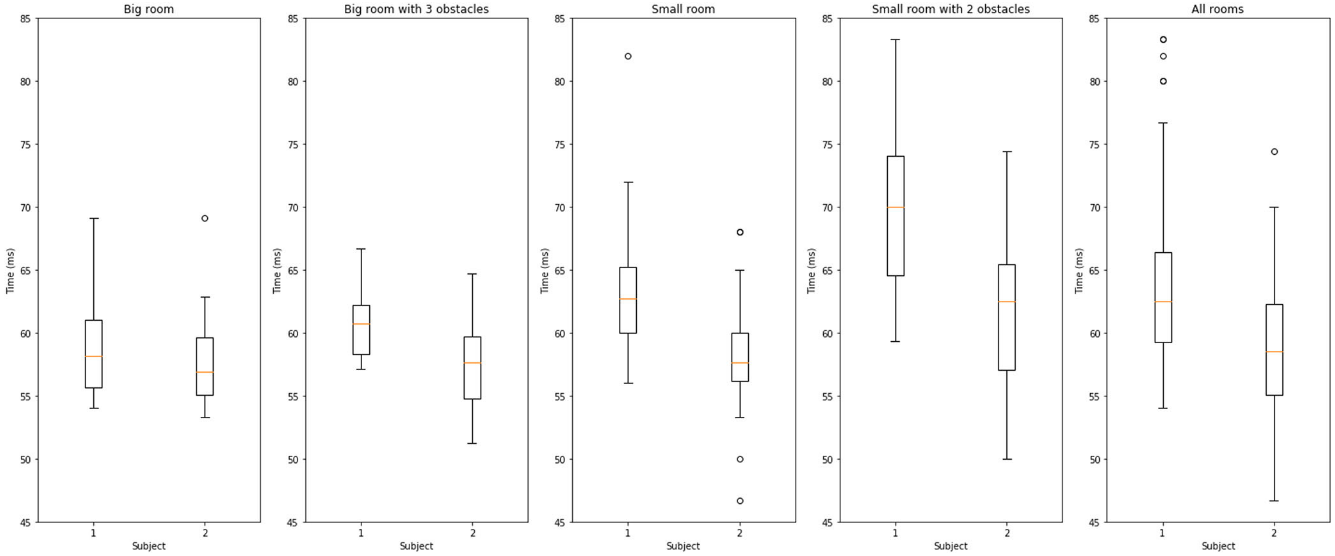

Fig. 7 Stride Time Variability. Stride time variability arising from the home environment. The panels from left to right show box plots of passively recorded stride time for two subjects in four different settings and in aggregate. The error bars of boxplot are generated by matplotlib using the matplotlib.pyplot.boxplot function with default parameters. The boxplot extends from the first to the third quartile of the data with a line at the median. The whiskers extend from the box by 1.5 times the inter-quartile range.

lower variability, due to the lack of sufficient motor performance variability during measurement periods[34]. A pragmatic approach is to view active assessment as more suitable to the measurement of subject capacity and passive assessment as a mechanism to capture real-life ability[35]. The above observation does not preclude adopting a hybrid approach if necessary. In this case, the presented framework and case studies still offers a useful guide to determine the potential influence of SoVs on study-specific Concepts of Interest.

Further, a key motivation in initiating this work was the need to clearly contrast variability due to the measurement process against variability caused by the disease. To this end, we believe that a core requirement towards the further development of mitigation techniques for a wider range of SoVs is to place greater emphasis on normative data sets reflecting performance by healthy subjects. This information is critical to establish ground truths of expected variability.

Finally, an inherent characteristic of DHTs is the rapid rate of advance in sensor technologies and the ability of modern software tools, such as machine learning and artificial intelligence, to improve their quantitative performance. Such rapid innovation can exacerbate the impact of SoVs, for example hardware used in a prospective clinical study might become outdated by the time the study is finished; or, algorithm performance might be enhanced by updating the software mid-study based on additional training data. Clearly, SoVs introduced by the availability of improved tools must also be managed in adopting a similar approach to the suggested SOV framework presented here. Alternatively, requiring new prospective studies for every major hardware, firmware, or model upgrade would represent a major barrier to innovation.

METHODS

3DT working group on SoVs

Created in partnership with Parkinson's UK, the Critical Path for Parkinson's Consortium (CPP) is a global initiative supporting collaboration among scientists from the biopharmaceutical industry, academia, government agencies, and patient-advocacy associations. The value of such collaborations is recognized by global regulatory agencies, including the US Food and Drug Administration and the European Medicines Agency, which have actively encouraged data-driven engagement through multi-stakeholder consortia[36]. A foundational tenet of CPP is the precompetitive collaborative nature of the consortium that forms the core principle for advancing the regulatory maturity of DHTs, and thus, facilitate their use in future clinical trials. To this end, CPP established the Digital Drug Development Tool (3DT) project, a precompetitive collaboration, aiming to align knowledge, expertise, and data sharing of DHTs across its consortium. Its main goal is to complement standard clinical assessments with a set of candidate objective digital measures, which can provide high precision measurements of disease progression and response to treatment.

This paper reports on the findings of the CPP 3DT: Sources of Variability (SoVs) Working Group. To develop the conceptual framework for the identification and characterisation of SoVs presented here, the WG adopted a triangulation methodology incorporating findings reported in the current research literature, direct experience with proprietary or unpublished work contributed by individual WG members, and data-driven analysis of key cases studies identified.

Data sets

Data used in Case Study 1 were pilot data obtained from Wearable Assessments in the Clinic and Home in PD (WATCH-PD), a 12-month multicentre, longitudinal, digital assessment study of PD progression in subjects with early untreated PD (clinicaltrials.gov#: NCT03681015). Its primary goal is to generate and optimize a set of candidate objective digital measures to complement standard clinical assessments in measuring the progression of disease and the response to treatment. A secondary goal is to understand the relationship between standard clinical assessments, research- grade digital tools used in a clinical setting, and more user-friendly consumer digital platforms to develop a scalable approach for objective, sensitive, and frequent collection of motor and nonmotor data in early PD. The clinical protocol[28] includes: (a) in-clinic assessments using six APDM Opal inertial measurement unit (IMU) sensors[37] that are placed in the lumbar region, sternum, wrists, and feet of the subject, which record accelerometer and gyroscope signals during a series of mobility-related tasks; and (b) a walking task performed at-home, where patients are instructed to place an iPhone in a pouch provided, and attach it to the lower back, and then initiate sensor data recording using an Apple Watch. The WATCH-PD trial has been approved by the WIRB Copernicus Group (protocol code WPD-01 and date of approval 12/21/2018). Informed consent was obtained from all subjects involved in the study. Written consent will not be obtained from participating participants since they are not identifiable by the study team. Participants are only identifiable at the study site level.

The data set used in Case Study 2 was obtained during software testing (quality improvement and usability) by the Digital Outcome Measure Variability due to Environmental Context Differences using Wearables project (DOMVar), conducted collaboratively between Birkbeck College, University of London, University College London and Panoramic Digital Health (who provided the study device cf. https://www.panoramicdigitalhealth.com/). The project was conducted according to The European Code of Conduct for Research Integrity (2017) and the guidelines of the Code of Practice on Research Integrity of Birkbeck College, University of London, and approved by the Ethics Committee of Birkbeck College, University of London. Informed consent was obtained from all subjects involved in the study.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Software quality testing data used in Case Study 2 are available to qualified researchers from co-author DH. Data presented in in Case Study 1 is from the ongoing WATCH-PD study and cannot be shared until completion and dissemination of results and approval by study sponsor. This is expected to become possible within 24 months from the acceptance date of this paper. Qualified researchers will be able to contact co-author RD at the University of Rochester, to request access to the data.

CODE AVAILABILITY

Python source code of the custom software used to conduct the calculations presented in Case Studies 1 and 2 is available upon reasonable request from co-authors KPK and DH correspondingly. In Case Studies 1, the APDM Mobility Lab software (cf. https://apdm.com/mobility/) now part of Clario, was used with default parameters. Co-author TRH developed the custom software employed using python and incorporating the GaitPy module. GaitPy was implemented following the parameters suggested by ref. [31].

Received: 21 October 2021; Accepted: 29 June 2022;
Published online: 15 July 2022

REFERENCES

1. Schneider, R. B. et al. Remote administration of the MDS-UPDRS in the time of COVID-19 and beyond. *J. Parkinsons Dis.* **10**, 1379–1382 (2020).
2. Arora, S. et al. Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD. *Neurology* **91**, e1528–e1538 (2018).
3. Bot, B. M. et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci. Data* **3**, 160011 (2016).
4. Stamate, C. et al. The cloudUPDRS app: A medical device for the clinical assessment of Parkinson's Disease. *Pervasive Mob. Comput.* **43**, 146–166 (2018).
5. Warmerdam, E. et al. Long-term unsupervised mobility assessment in movement disorders. *Lancet Neurol.* **19**, 462–470 (2020).
6. Sacks, L. & Kunkoski, E. Digital health technology to measure drug efficacy in clinical trials for parkinson's disease: a regulatory perspective. *J. Parkinsons Dis.* **11**, S111–S115 (2021).
7. Jha, A. et al. The CloudUPDRS smartphone software in Parkinson's study: cross-validation against blinded human raters. *npj Parkinson's Dis.* **6**, 1–8 (2020).
8. Mei, J., Desrosiers, C. & Frasnelli, J. Machine learning for the diagnosis of Parkinson's disease: a review of literature. *Front Aging Neurosci.* **13**, 633752 (2021).

9. Zhan, A. et al. Using smartphones and machine learning to quantify parkinson disease severity: the mobile parkinson disease score. *JAMA Neurol.* **75**, 876–880 (2018).

10. Taylor, K. I., Staunton, H., Lipsmeier, F., Nobbs, D. & Lindemann, M. Outcome measures based on digital health technology sensor data: data- and patient-centric approaches. *NPJ Digital Med.* **3**, 1–8 (2020).

11. Arora, S. et al. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. *Parkinsonism Relat. Disord.* **21**, 650–653 (2015).

12. Goetz, C. G. et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov. Disord.* **23**, 2129–2170 (2008).

13. Narayana, R., Hellman, B., Addyman, C. & Stamford, J. Self-management in long term conditions using smartphones: A pilot study in Parkinson's disease. *Int. J. Integrated Care* **14**, (2014). https://www.ijic.org/articles/abstract/10.5334/ijic.1820/#.

14. Bettini, C. et al. A survey of context modelling and reasoning techniques. *Pervasive Mob. Comput.* **6**, 161–180 (2010).

15. Servia-Rodríguez, S. et al. Mobile Sensing at the Service of Mental Well-being: a Large-scale Longitudinal Study. In *Proceedings of the 26th International Conference on World Wide Web* 103–112 (International World Wide Web Conferences Steering Committee, 2017). https://doi.org/10.1145/3038912.3052618.

16. Stephenson, D. et al. Precompetitive consensus building to facilitate the use of digital health technologies to support Parkinson disease drug development through regulatory. *Sci. DIB* **4**, 28–49 (2020).

17. Hill, D. L. et al. Metadata framework to support deployment of digital health technologies in clinical trials in Parkinson's disease. *Sens. (Basel)* **22**, 2136 (2022).

18. Zaiyadi, N., Mohd-Yasin, F., Nagel, D. & Korman, C. Reliability measurement of single axis capacitive accelerometers employing mechanical, thermal and acoustic stresses. in 1–2 (2010). https://doi.org/10.1109/ISDRS.2009.5378027.

19. Badawy, R. et al. Automated quality control for sensor based symptom measurement performed outside the lab. *Sens. (Basel)* **18**, E1215 (2018).

20. Grillo, E. U., Brosious, J. N., Sorrell, S. L. & Anand, S. Influence of smartphones and software on acoustic voice measures. *Int J. Telerehabil* **8**, 9–14 (2016).

21. Tsanas, A., Little, M. A., McSharry, P. E. & Ramig, L. O. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Trans. Biomed. Eng.* **57**, 884–893 (2010).

22. Ben Mansour, K., Rezzoug, N. & Gorce, P. Analysis of several methods and inertial sensors locations to assess gait parameters in able-bodied subjects. *Gait Posture* **42**, 409–414 (2015).

23. Fasel, B. et al. A wrist sensor and algorithm to determine instantaneous walking cadence and speed in daily life walking. *Med Biol. Eng. Comput* **55**, 1773–1785 (2017).

24. Lim, H., Kim, B. & Park, S. Prediction of lower limb kinetics and kinematics during walking by a single IMU on the lower back using machine learning. *Sensors (Basel)* **20**, (2019).

25. Yurtman, A. & Barshan, B. Activity recognition invariant to sensor orientation with wearable motion. *Sens. Sens.* **17**, 1838 (2017).

26. Sabatini, A. M. Quaternion-based extended Kalman filter for determining orientation by inertial and magnetic sensing. *IEEE Trans. Biomed. Eng.* **53**, 1346–1356 (2006).

27. Derungs, A. & Amft, O. Estimating wearable motion sensor performance from personal biomechanical models and sensor data synthesis. *Sci. Rep.* **10**, 11450 (2020).

28. Adams, J. L. WATCH-PD: Wearable assessments in the clinic and home in Parkinson's disease: study design and update. *Mov. Disord.* **35**, S1–S702 (2020).

29. El-Gohary, M. et al. Continuous monitoring of turning in patients with movement disability. *Sens. (Basel)* **14**, 356–369 (2013).

30. Czech, M. & Patel, S. GaitPy: an open-source python package for gait analysis using an accelerometer on the lower back. *J. Open Source Softw.* **4**, 1778 (2019).

31. Del Din, S., Godfrey, A. & Rochester, L. Validation of an accelerometer to quantify a comprehensive battery of gait characteristics in healthy older adults and parkinson's disease: toward clinical and at home use. *IEEE J. Biomed. Health Inf.* **20**, 838–847 (2016).

32. Zhou, L. et al. How we found our IMU: Guidelines to IMU selection and a comparison of seven imus for pervasive healthcare applications. *Sensors (Basel)* **20**, 4090 (2020).

33. Perraudin, C. G. M. et al. Observational study of a wearable sensor and smartphone application supporting unsupervised exercises to assess pain and stiffness. *Digit Biomark.* **2**, 106–125 (2018).

34. Giladi, N., Horak, F. B. & Hausdorff, J. M. Classification of gait disturbances: distinguishing between continuous and episodic changes. *Mov. Disord.* **28**, 1469–1473 (2013).

35. Atrsaei, A. et al. Gait speed in clinical and daily living assessments in Parkinson's disease patients: performance versus capacity. *NPJ Parkinsons Dis.* **7**, 24 (2021).

36. Maxfield, K. E., Buckman-Garner, S. & Parekh, A. The role of public-private partnerships in catalyzing the critical path. *Clin. Transl. Sci.* **10**, 431–442 (2017).

37. Mancini, M. & Horak, F. B. Potential of APDM mobility lab for the monitoring of the progression of Parkinson's disease. *Expert Rev. Med. Devices* **13**, 455–462 (2016).

ACKNOWLEDGEMENTS

The authors thank Michael Lawton for support in preparation of the manuscript. Critical Path Institute is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) and is 56.5% funded by FDA/HHS, totaling $16,749,891, and 43.5% funded by non-government source(s), totaling $12,895,366. 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. For more information, please visit FDA.gov. The Critical Path Institute's Critical Path for Parkinson's (CPP) 3DT initiative consists of a subset of CPP member organizations and is funded by following members: AbbVie, Biogen, GSK, Lundbeck, Merck Sharp and Dohme, Roche, Takeda, UCB, and the Michael J Fox Foundation. The 3DT team recognizes leading academic advisors, Jesse Cedarbaum, Glenn Stebbins, Tanya Simuni, Bas Bloem, Marjan Meinders, Melissa Kostrzebski, Parkinson's UK, and people living with Parkinson's for their role in contributing to this multinational collaborative initiative.

AUTHOR CONTRIBUTIONS

G.R.: Conception and paper drafting. T.R.H.: Conception, data analysis (Case Study 1) and paper drafting. D.H.: Conception, data analysis (Case Study 2) and paper drafting. A.V.D.: Substantial revisions. M.M.: Substantial revisions. L.J.W.E.: Substantial revisions. J.B.: Substantial revisions. A.D.: Substantial revisions. K.F.: Substantial revisions. K.P.K.: Additional data analysis (Case Study 1) and substantial revisions. N.M.: Substantial revisions. R.B.: Substantial revisions. S.S.: Substantial revisions. D.S.: Conception and paper drafting. J.A.: Conception of WATCH-PD study, data collection and substantial revisions. E.R.D.: Conception of WATCH-PD study, data collection and substantial revisions. J.C.: Substantial revisions. All authors have read and approved the manuscript.

COMPETING INTERESTS

D.S., M.M., R.B., S.S., J.B. (at time of writing at C-Path): No competing interests to declare. G.R.: is a technical advisor to the Critical Path for Parkinson's Consortium (CPP). T.R.H.: at time of writing Project Lead of Digital and Quantitative Medicine at Biogen. D.H.: is President of Panoramic Digital Health. A.V.D.: is Director of Digital Strategy at Takeda. J.C.: is Director of Digital Health Strategy at AbbVie and Industry Co-Director of CPP. L.E.: has received funding from the Michael J Fox Foundation for Parkinson's Research, UCB Biopharma, Stichting ParkinsonFonds, the Dutch Ministry of Economic Affairs, and ZonMw. A.D.: is Digital Biomarker Sensor Lead & Senior Scientist at Roche. K.F.: at time of writing Senior Director of Digital and Quantitative Medicine at Biogen. K.P.K.: Scientist at Biogen. N.M.: at time of writing Senior Principal Scientist of Quantitative Pharmacology and Pharmacometrics at Merck. R.D.: has stock ownership in Grand Rounds, an online second opinion service, has received consultancy fees from 23andMe, Abbott, Abbvie, Amwell, Biogen, Clintrex, CuraSen, DeciBio, Denali Therapeutics, GlaxoSmithKline, Grand Rounds, Huntington Study Group, Informa Pharma Consulting, medical-legal services, Mednick Associates, Medopad, Olson Research Group, Origent Data Sciences, Inc., Pear Therapeutics, Prilenia, Roche, Sanofi, Shire, Spark Therapeutics, Sunovion Pharmaceuticals, Voyager Therapeutics, ZS Consulting, honoraria from Alzheimer's Drug Discovery Foundation, American Academy of Neurology, American Neurological Association, California Pacific Medical Center, Excellus BlueCross BlueShield, Food and Drug Administration, MCM Education, The Michael J Fox Foundation, Stanford University, UC Irvine, University of Michigan, and research funding from Abbvie, Acadia Pharmaceuticals, AMC Health, BioSensics, Burroughs Wellcome Fund, Greater Rochester Health Foundation, Huntington Study Group, Michael J. Fox Foundation, National Institutes of Health, Nuredis, Inc., Patient-Centered Outcomes Research Institute, Pfizer, Photopharmics, Roche, Safra Foundation. J.A.: has received honoraria from Huntington Study Group, research support from National Institutes of Health, The Michael J Fox Foundation, Biogen, Safra Foundation, Empire Clinical Research Investigator Program, and consultancy fees from VisualDx and Spark Therapeutics.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00643-4.

Correspondence and requests for materials should be addressed to George Roussos.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.