












Remote smartphone monitoring of Parkinson's disease and individual response to therapy

Larsson Omberg^{1,10}  , Elias Chaibub Neto^{1,10}  , Thanneer M. Perumal¹, Abhishek Pratap^{1,2}, Aryton Tediario¹, Jamie Adams^{3,4}, Bastiaan R. Bloem⁵, Brian M. Bot¹, Molly Elson³ , Samuel M. Goldman⁶, Michael R. Kellen¹, Karl Kieburz^{3,4}, Arno Klein¹, Max A. Little^{7,8}, Ruth Schneider^{3,4}, Christine Suver¹ , Christopher Tarolli^{3,4}, Caroline M. Tanner⁶, Andrew D. Trister⁹ , John Wilbanks¹, E. Ray Dorsey^{3,4}  and Lara M. Mangravite¹ 

Remote health assessments that gather real-world data (RWD) outside clinic settings require a clear understanding of appropriate methods for data collection, quality assessment, analysis and interpretation. Here we examine the performance and limitations of smartphones in collecting RWD in the remote mPower observational study of Parkinson's disease (PD). Within the first 6 months of study commencement, 960 participants had enrolled and performed at least five self-administered active PD symptom assessments (speeded tapping, gait/balance, phonation or memory). Task performance, especially speeded tapping, was predictive of self-reported PD status (area under the receiver operating characteristic curve (AUC) = 0.8) and correlated with in-clinic evaluation of disease severity ($r = 0.71$; $P < 1.8 \times 10^{-6}$) when compared with motor Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). Although remote assessment requires careful consideration for accurate interpretation of RWD, our results support the use of smartphones and wearables in objective and personalized disease assessments.

RWD offers the opportunity to improve our understanding and management of health and disease outside the clinical setting¹. An increasingly popular method for collecting RWD is the use of remote digital assessments that allow frequent sampling and have been used to aid in the diagnosis, treatment and monitoring of several conditions, including atrial fibrillation and diabetes^{2,3}. When used in population-based studies, remote assessment can increase understanding of heterogeneity in disease manifestation, the effect of disease on daily living and quality of life and the effect of non-medical factors on health⁴. Frequent longitudinal assessments can also serve as an essential tool for guiding personalized care management. Furthermore, in the clinical trial setting, there is the possibility that objective remote assessments may produce improved outcome measures with increased sensitivity to detect early changes in health. However, the complexities in collection and analysis^{5–7} of meaningful real-world evidence presents clear obstacles.

Remote digital studies can shift the user experience compared with traditional clinical studies and generate new challenges for the researcher that must be addressed to run an effective study. For participants, digital studies often require frequent engagement in the midst of their daily routines, unassisted interpretation of instructions, adherence to data collection protocols and a willingness to share information gathered through their mobile devices. Researchers must develop dynamic communication and engage-

ment strategies to keep participants motivated. Failure to address these emerging needs have led to engagement and retention issues^{8,9}. Furthermore, hidden biases, confounds and autocorrelation in longitudinally collected data without supervision require analytical approaches that account for these issues. For these reasons, methodologies for effective collection, analysis and interpretation of digital measures are under active investigation.

In recent years, there has been an emphasis on the use of sensors to develop objective assessments for conditions that are traditionally subjectively evaluated, such as neuropsychiatric or neurodegenerative disorders^{10–15}. PD is especially well suited for sensor-based assessments because of the prominent motor symptoms in PD^{14,15}. Motor symptoms include slow movement, impaired gait, diminished balance, soft voice and tremor. Several sensor-based devices have been used in clinical research^{16–18}. While this can improve clinical assessment, many of these devices are not widely accessible or tolerated outside the clinical setting and do not capture the frequent fluctuations in symptoms experienced over time by many patients with PD. Consumer-focused devices such as smartphones offer an alternative that is widely available to patients and have the potential to provide objective, frequent and sensitive assessments using their embedded high-fidelity sensors⁸. Several smaller clinical studies indicate that smartphone sensors with the right protocol have the capability to capture biologically meaningful digital assessments

¹Sage Bionetworks, Seattle, WA, USA. ²Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA.

³Center for Health and Technology, University of Rochester Medical Center, Rochester, NY, USA. ⁴Department of Neurology, University of Rochester Medical Center, Rochester, NY, USA. ⁵Radboud University Medical Center; Donders Institute for Brain, Cognition and Behaviour; Department of Neurology, Nijmegen, the Netherlands. ⁶Department of Neurology, University of California–San Francisco and Parkinson's Disease Research, Education and Clinical Center, San Francisco Veterans Affairs Health Care System, San Francisco, CA, USA. ⁷School of Computer Science, University of Birmingham, Birmingham, UK. ⁸Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁹Bill & Melinda Gates Foundation, Seattle, WA, USA. ¹⁰These authors contributed equally: Larsson Omberg, Elias Chaibub Neto. ✉e-mail: larsson.omberg@sagebionetworks.org; elias.chaibub.neto@sagebionetworks.org; ara.mangravite@sagebionetworks.org

of PD^{10,13,19–23}. Given the dynamic and heterogeneous nature of PD and the availability of viable digital assessments, remote monitoring of PD provides a powerful opportunity to evaluate the benefits and challenges of collecting and interpreting sensor data for real-world evidence.

In 2015, we initiated the mPower study to evaluate the possibilities and limitations of using mobile phones to collect real-world evidence across a diverse population in the context of a fully remote research study (study reviewed by Western IRB, see Methods). mPower collected smartphone-based sensor data from participants while they performed self-guided versions of assessments used in the clinic to evaluate motor symptoms of PD. A main objective of the mPower study was to support rapid advancement in the understanding and use of sensor data for health assessments. To this aim, participants were given the option of having analysis performed solely by the mPower study team or of having their data shared broadly with qualified researchers²⁴. Here we analyze data from participants that opted in and participants that opted out of broad data sharing. In addition, we report on a validation study, objectivePD (reviewed by Rochester IRB) in which 44 individuals were followed over 6 months including three in-clinic assessments in addition to remote smartphone-based assessments. We use this analysis to illustrate the feasibility and caveats of mobile health studies.

Results

The mPower study focused on the interaction between mobile health measures and the lived experience with PD and offers a case study of a remote mobile health study. Here we summarize the results of the study from several angles, including the following: recruitment and retention, the strong effects of confounders and bias from data collected in the real world, the opportunities for personalized analysis afforded by the high number of repeat measures from individuals over time and accuracy of self-report and validation compared with in-clinic measures.

Recruitment and study assessment. Over a 6-month period, from 9 March 2015 to 9 September 2015, 12,703 individuals from all 50 US states enrolled in the mPower research study. Enrollment was defined as downloading the smartphone application, completing the electronic informed consent, the enrollment survey and at least one activity (Fig. 1 and Table 1). Recruitment was solely by word of mouth and press coverage that was aided by attention to the initial launch of ResearchKit by Apple. Several thousand participants enrolled in the days after launch. Following this initial burst, a relatively constant rate of enrollment was maintained (60 ± 35 per week).

During the study period, participants were encouraged, but not obligated, to complete activities up to three times daily. These activities included a series of five active assessments (speeded tapping, voice, walk, balance and memory) during which sensors in the phone were used to collect data from participants. They were also asked to complete several patient-reported outcomes (PRO) surveys. Sensor data collected included detailed time series of touchscreen interactions during the tapping and memory assessment, audio recordings during the voice assessment and 100 Hz accelerometer and gyroscope measurements during the walk and balance assessments. Over 350,000 activities were performed across all participants in this 6-month time interval. Active assessments were performed nearly two orders of magnitude more frequently than survey questionnaires.

Participant engagement and retention. Of the 12,703 participants, 1,414 self-reported as having a professional diagnosis of PD and 8,432 self-reported not having a PD diagnosis (non-PD). On average, the PD participant population was significantly older, more educated and had a larger proportion of females compared to non-PD participants (Table 1). Furthermore, they were more

active in the study, contributing on average five to seven times more measurements per assessment. Long-term engagement followed a typical (for app use) exponential decay with a small number of individuals contributing thousands of activities and thousands of individuals contributing 1 or 2 activities^{8,9} (Supplementary Fig. 1). This has the unfortunate consequence that, for any analysis requiring long-term longitudinal data, the available sample size is considerably smaller than total enrollment (Fig. 1).

During enrollment, participants were given agency to decide whether their data should be retained by study organizers or should be shared broadly with qualified researchers²⁴; 67% of participants opted to share with all qualified researchers, and these data have been made available to the research community, whereas a further 3,183 chose to share only with the mPower study teams. Individuals who opted to share their data broadly were more likely to have PD, as determined by self-report of professional diagnosis, ($P < 0.0001$) and, on average, completed more active tapping, voice and walk/balance assessments (all $P < 0.0001$). All analyses described below were performed on the full study population; similar results were observed when conducted on the subpopulation that opted for broad data sharing²⁵.

Population-level analysis of self-administered health assessments. Because diagnostics are of principal interest for remote monitoring, we first evaluated whether cross-sectional analysis of these data could be used to develop a population-level classifier to distinguish PD from non-PD participants based on remote assessment performance. Diagnostic classifiers built from machine learning applications often fail to generalize because they do not account for biological and/or technical confounders present in the training data that are unrelated to disease state. Collection of sensor-based data from assessment performance in real-world studies is anticipated to introduce several sources of confound or bias, which would need to be addressed in the context of classification. Indeed, many of the factors that make remote research so compelling—the ability to enroll a diverse population—must be considered carefully in population-level assessments.

We first considered enrollment bias. In mPower, as with many remote studies, the research team did not control enrollment parameters beyond a set of minimal inclusion/exclusion criteria. For this reason, the study population may be unbalanced across factors that affect assessment performance but are independent of disease state. Some of these factors may be known and some may not. As mentioned above, the non-PD participants were younger, less educated and had a greater percentage of males compared with PD participants. To assess the influence of these potential sources of confounding in the mPower data, we used confounding tests²⁶ (Methods) to quantify the contribution of the education, gender and age variables to the predictive performance of the PD versus non-PD classification. The classification performance tended to be artificially boosted by the age confounder in all four assessments, and moderately inflated by the gender confounder in the voice (and walk and tapping assessments to a lesser degree) (Supplementary Note 1). Based on these results, age- and gender-matched participants were used in the mPower classification analysis to avoid age and gender confounding (Methods). It is nonetheless important to point out that we can only evaluate observed confounders, and that some unobserved confounders might still be biasing these results.

A second main source of potential confounding on classification analyses that use data frequently sampled from individuals, such as occurs in mobile health data, is the handling of repeated measures. When repeated measures collected from an individual are split across training and test sets (called record-wise splits)^{7,27}, classifiers are inadvertently performing participant identification rather than (or in addition to) the identification of biological traits such as disease status^{6,28}. To quantify the effect of this identity confounding

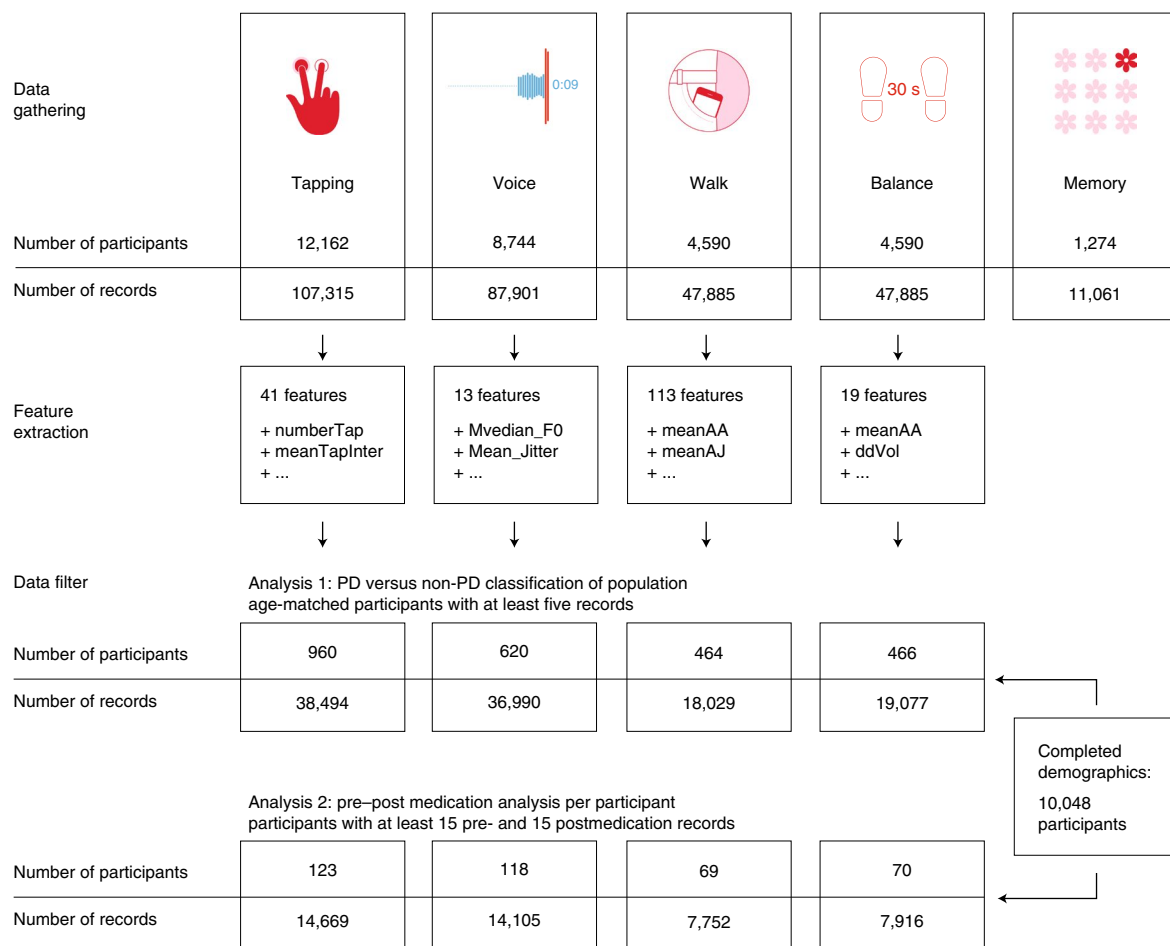


Fig. 1 | Data collection, processing and analysis of active performance assessments. A total of 12,703 individuals consented and participated in the first 6 months of the study. Individuals were encouraged but not obligated to complete all assessments. Level of participation varied drastically across individuals (Supplementary Fig. 7). For data gathering, we list the number of participants that performed the assessment at least once (with total number of assessments or records recorded below). Because the memory assessment seemed to reflect motor function rather than cognition, it was excluded from subsequent analysis. Raw sensor data was summarized into 13–113 features per assessment (Methods). Population-level analyses were performed using an age-matched subset of PD and non-PD participants who had provided at least five measures per assessment to make sure we captured participant variability (high users). Personalized analyses were performed using the subset of PD participants who provided 15 measures before and 15 measures after self-reported administration of PD medication.

on classification performance, we used subject-wise permutations on record-wise splits of the mPower data^{6,28}. Indeed, the mPower data were highly vulnerable to this issue of identity confounding (Supplementary Fig. 2) with record-wise splits leading to overly optimistic assessments of classification performance.

To control for this issue of identity confounding in the mPower classification analysis, age-/gender-matched classifiers were built using collapsed summaries of longitudinal data or ‘subject-wise’ splits of longitudinal data across the training and test sets (Methods). These PD classifiers (Supplementary Figs 3 and 4) had lower performance than those developed using repeat measure and records-wise splits (Supplementary Fig. 2), but better reflect the real ability of the classifier to separate PD and non-PD participants in a new sample based on the disease characteristics and in accordance with previous results in the literature²⁹.

Classification performance varied across assessments. Using a subset of participants who have contributed at least five repeated assessments (with sample sizes per assessment ranging from 464 to 960 as shown in Fig. 1), a random forest classifier achieved AUC = 0.80 ($P < 10^{-31}$) for PD classification using collapsed features

from the tapping assessment, where the P value is testing if the classification performance was better than random (Methods). The performances using collapsed features from each of the other activities were more modest (AUCs equal to 0.62 ($P < 10^{-4}$) for walk, 0.65 ($P < 10^{-5}$) for balance and 0.60 ($P < 10^{-4}$) for voice).

Variation in activity performance across repeated measures.

Activity performance tended to be more variable in PD participants than in non-PD participants. By looking at day-to-day fluctuations across all individuals who participated for at least 15 distinct days, 17 out of 186 features showed significantly greater intraindividual interquartile range (false discovery rate (FDR) < 0.05) in PD cases compared with only 4 out of 186 features with significantly greater variability in non-PD cases (Supplementary Figs. 5 and 6). Qualitative interviews with participants have indicated that some PD participants were interested in remaining active in this study precisely because they were interested in monitoring this variability in performance over time³⁰.

Personalized medication response analysis. Most PD patients are treated with medications to alleviate disease symptoms, with

Table 1 | Baseline characteristics of mPower and ObjectivePD study populations

mPower	Total population			ObjectivePD population		
	PD	Non-PD	P value	PD	Non-PD	P value
N	1,414	8,432	–	21	23	–
Age	60	32	<10 ^{–16}	64	44	4 × 10 ^{–4}
Sex (women (%))	34	19	<10 ^{–36}	52	59	0.9
Race (White (%))	88	68	9 × 10 ^{–4}	100	95	0.1
Education (completed college (%))	70	56	0.06	65	39	4.3 × 10 ^{–4}
PD duration (years)	5.3	–	–	3.1	–	–
Proportion taking levodopa (%)	96	–	–	62	–	–
MDS-UPDRS total score	–	–	–	38.4	8.2	9.5 × 10 ^{–9}

levodopa as the standard treatment. For many patients, treatment effectively reduces symptoms and the patient experiences an ‘on’ state throughout the day. For others, treatment is less effective and the patient experiences fluctuations in symptom severity over the course of each day (‘on/off’ fluctuations). Because many participants take their medication at the same time every day, evaluation of medication effects on performance collected in this manner may be confounded with diurnal factors relating to the time of day (ToD) that the activity was performed. Such factors might be caused by circadian rhythms or by daily routine, such as coffee intake in the morning, in addition to a medication effect. To address this, we evaluated whether variation in performance observed in PD participants reflects medication and/or other temporal effects. Because activity performance was highly individualized and heterogeneous across the population, we developed personalized methods to monitor changes in disease severity due to the effects of medication or other disease modulators over time. The longitudinal data from each individual were analyzed separately using a set of conditional independence (CI) tests to disentangle potential medication effects from ToD effects²⁵. Because of the potential presence of temporal autocorrelation in the longitudinal data, these CI tests were implemented using temporal regression models based on robust estimators of standard errors that can account for heteroscedasticity and autocorrelation in the data^{31,32}. Looking at data from participants that had contributed at least 15 activities before medication and at least 15 after (with number of participants per task ranging from 69 to 123 (Fig. 1) and number of records per participant ranging from 31 to 448 (Supplementary Fig. 7)), we identified that 47% of PD participants experienced fluctuations that could possibly be attributed to medication (at a FDR < 0.05). The effect size based on relative importance³³ (Methods) among the fluctuators was, on average, 0.17 (Fig. 2) and the direction of the effect for 71% of the features was consistent with the directions between PD and non-PD. Given that some symptoms (for example, dyskinesia) can be induced by medication, this is not inconsistent with expectation. The significance of these fluctuations as determined by the *P* values was significantly associated with the disease onset year of the participant (*P* < 0.001), recapitulating clinical observations that participants with longer PD history tend to show stronger medication-mediated on/off fluctuations in performance (Fig. 2 and Supplementary Fig. 8).

Performance as measured by assessment features varied substantially relative to ToD, indicating a potential effect from diurnal variation (Fig. 2d and Methods). These effects were not associated with the year of disease onset (Fig. 2e; *P* > 0.85). The distribution of features associated with change in medication state varied across individuals and across assessments: 35% of features for tapping, 18% for voice, 24% for walk and 8% for balance. Changes in performance were highly individualized, with strong on/off fluctuations observed in different assessments across individuals and often in

one activity but not in another (Supplementary Fig. 9). This observation recapitulates the known variation in individual patient-level PD symptomatology³⁴.

Consistency of survey answers. Although it is not possible to evaluate the accuracy of data collected in an unsupervised manner, we sought to evaluate the consistency of data across repeat measures using responses to the demographic survey. To this aim, we calculated the proportion of participants who provided inconsistent answers across repeated administrations of survey questions. This included nonunique answers to questions such as the year the participant was diagnosed, the year the participant started taking medications, the year of the disease onset and the race of the participant. In all cases, the inconsistency rates were lower than 0.64%. (The rates for the respective questions were, respectively, 0.24%, 0.31%, 0.64% and 0.55%.)

Correlation of smartphone performance measures with in-clinic PD assessments. Using a subcohort recruited at the University of Rochester comprising 44 individuals (21 with clinically diagnosed PD and 23 non-PD controls) as part of the ObjectivePD substudy (Table 1) we compared mPower symptom severity scores with in-clinic PD severity measures. The mPower symptom severity score was derived as the average of the prediction probabilities from each of the classifiers for the PD versus non-PD (one for each of the smartphone assessments, tapping, walk, balance and voice). This model, albeit simpler, mirrors similar efforts in the literature^{22,23}. This mPower severity score was associated with total MDS-UPDRS (Pearson *r* = 0.71, Bonferroni corrected *P* < 1.8 × 10^{–6}), Schwab and England Activities of Daily Living Scale (SE-ADL) (*r* = –0.61, Bonferroni corrected *P* < 2.3 × 10^{–5}) and Hoehn and Yahr score (Bonferroni corrected *P* < 3 × 10^{–8}) (Fig. 3).

Discussion

Digital technologies provide the potential to collect remote, objective, real-world assessments that may meaningfully affect both clinical research and clinical care. Appropriate implementation of these approaches provides an opportunity to study health in the context of daily life. However, inappropriate implementation can lead to unsatisfactory results or incorrect conclusions. To evaluate the feasibility and limitations of real-world studies using digital tools, we conducted a completely remote study that used smartphone sensor-based assessments to collect repeated measures with the goal of objectively quantifying PD.

One of the opportunities of remote research studies is to overcome some of the existing limitations in diversity across participant populations that are observed with in-clinic studies. Although the widespread availability of smartphones and other pervasive technologies has the potential to increase research accessibility across a

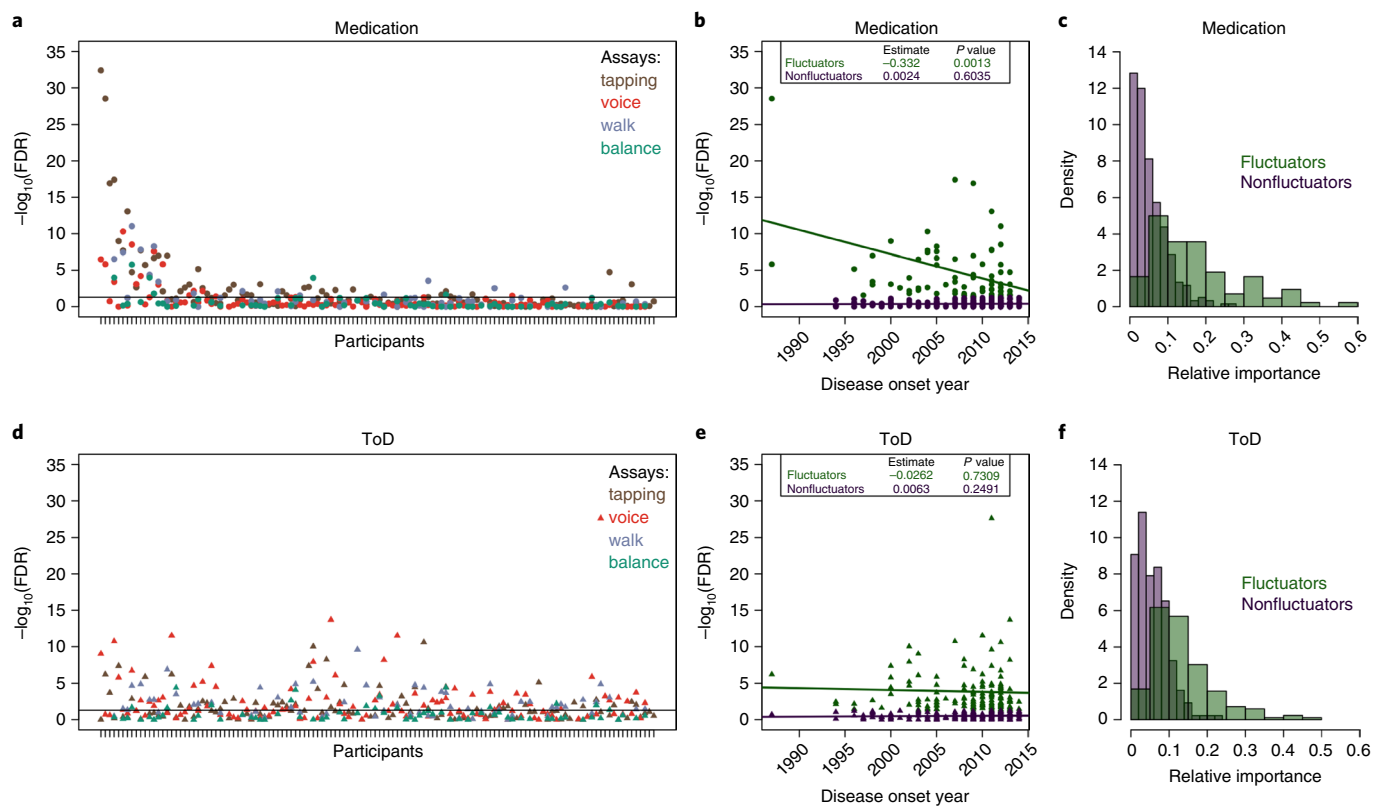


Fig. 2 | Personalized fluctuations in performance across assessment and association with on/off motor fluctuations and diurnal variations. a, FDR-adjusted P values (in $-\log_{10}$ scale) from the union-intersection (UI) tests for putative medication effects using Newey–West regression (Methods). The color-coded dots represent the four activity tasks. Each position in the x axis represents a participant, and reports the UI test P values for each of the four activity tasks. For each activity task, the UI test P value of each participant corresponds to the minimum FDR-corrected P value across all features of the participant. The y axis reports the ‘double corrected’ UI test P values, where, in addition to ‘within-participant’ correction across the features, we also perform a second round of corrections across the participants, correcting for each activity separately. The horizontal black line corresponds to a significance threshold of 0.05. **b,** Corrected P values versus disease onset year for each participant, showing that medication fluctuators ($\text{FDR} \leq 0.05$) tend to be associated with disease history. **c,** Distribution of the relative importance of medication, that is, the proportion of the variability that is explained by the medication, faceted by on/off medication fluctuators. **d,** FDR-adjusted P values (in $-\log_{10}$ scale) from the UI tests for putative ToD effects using Newey–West regression. Color-coded dots, positions in the x axis and activity tasks as shown in **a**. **e,** Corrected P values versus disease onset year for each participant, showing that ToD fluctuators ($\text{FDR} \leq 0.05$) are not associated with disease history. **f,** Distribution of the relative importance of ToD, that is, the proportion of the variability that is explained by ToD, faceted by on/off ToD fluctuators. (Supplementary Fig. 8 presents analogous results based on ARIMA and standard-regression models.)

wide range of demographic classes, this approach does not necessarily balance enrollment across groups. As with all observational studies, study design and recruitment protocols have a great effect on the study population. In mPower, recruitment supported self-selected enrollment of individuals who had an Apple smartphone, resided in the United States and spoke English. These constraints skewed study enrollment towards younger individuals with higher incomes and education levels. This was particularly true for the non-PD participants. The self-enrolled population of PD participants, who had a more distinct set of reasons for participation than the non-PD individuals, were significantly older than the self-enrolled population of non-PD participants. Because both age and disease status have an effect on performance on assessments, this led to a clear confounder that had to be addressed in the analysis. In any observational study, it is essential to assess the influence of observed confounders and to perform statistical adjustments (such as matching) if necessary. Such remedies, however, can only account for observed confounders, and provide no protection against unmeasured confounders. This is a principal concern with remote studies that may have limited information about their participants with which to identify potential confounders. Indeed, this is a problem observed

across any study that enrolls through self-selection, and may unintentionally affect analytical outcomes^{35,36}.

The ability to properly validate the accuracy of data collected in an unsupervised manner is a main concern with RWD analysis. First, there are questions about the reliability of self-reported data. Using a subset of individuals who changed responses to their demographic survey, we did observe within-individual consistency within 99% across repeat survey questions. Second, there is concern around identity confirmation where there is no assurance that all data collected in an unsupervised manner from a research participant is actually provided by the same individual. Although this may be an issue for low-dimensional data, such as survey questionnaires, sensor-based data are highly individualized. Indeed, our analysis of mPower data has demonstrated that machine learning approaches are able to identify patterns associated with individual participants across repeat measures. This observation indicates that (given enough longitudinal data) it might be possible to incorporate routine checks for identity confirmation in data collection that is stable to disease progression. Third, unsupervised assessment administration may result in inappropriate performance due to misunderstanding of instructions or environmental limitations.

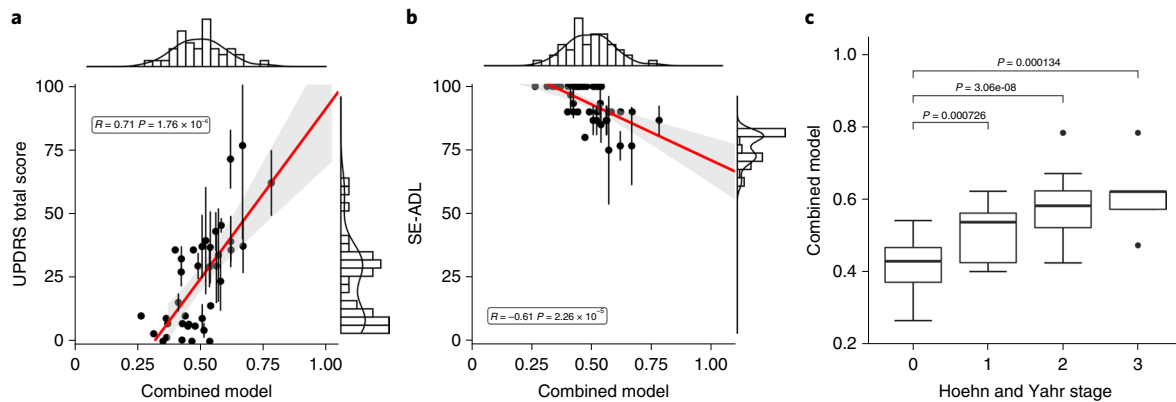


Fig. 3 | Association between mPower active assessments collected outside clinic and disease severity as measured in-clinic in the objectivePD cohort. **a–c.** Performance on active assessments was summarized into a combined model consisting of average performance across the four activity assessments (balance, tapping, walk and voice) as determined by random forest model (Supplementary Fig. 3). This combined model was correlated with total UPDRS (Bonferroni corrected $P = 1.8 \times 10^{-6}$, $r = 0.71$) (**a**), SE-ADL (corrected $P = 2.3 \times 10^{-5}$, $r = -0.61$) (**b**), and Hoehn and Yahr staging P values from a two-tailed t -test with Bonferroni correction (**c**). In **a** and **b**, error bars represent s.d. from repeated MDS-UPDRS measurements. In **c**, the distributions are represented as boxplots where the upper and lower bounds of the box represent third (Q3) and first (Q1) quartiles, respectively; the horizontal line is the median; the upper and lower whiskers are given by Q3 + 1.5 IQR and Q1 – 1.5 IQR, respectively; outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing the minimum and maximum, respectively. All results are based on a cohort of 44 participants.

This surfaced as a concern in mPower, as noted in other similar smartphone-based studies³⁷. A usability study of the mPower app (performed after the app was released, using a small sample of nine participants) indicated that a fair proportion of the participants failed to fully understand the instructions and, indeed, evidence for incorrect performance of assessments could be directly observed in the data (Supplementary Fig. 10). As an example, the walking assessment instructed participants to put the phone in the pocket and walk, if possible, in a straight line for 30 s. Inspection of the acceleration signals shows a full spectrum of noncompliant behaviors in a small subset of performances, ranging from removing the phone from the pocket during the activity to performing the assessment without carrying their phone. Such issues may be partially addressed through clear instructions, animations and/or video tutorials—all of which have been incorporated into later administrations of mPower (<https://parkinsonmpower.org/your-story>)—but must also be addressed in analysis through administration of quality control measures designed to disregard potentially faulty data³⁷.

RWD tracking provides a potential opportunity to monitor the effect on health assessments of both medical and nonmedical factors throughout the course of daily life. In clinical care, remote symptom tracking can be used to inform clinicians in managing patient treatment strategies, including drug dosing. In the short term, RWD from smart devices can replace patient recall to evaluate treatment efficacy between visits. In the longer term, these approaches may enable a continuous monitoring system that allows clinicians to manage patient care more dynamically and effectively. In clinical trials, there are also efficiencies that can be realized from rapid and real-time monitoring. These approaches allow an expanded assessment of drug effects to include participant-driven measures of treatment efficacy. In addition, the longitudinal nature of RWD streams enables trial designs that focus on precision-medicine approaches.

In both clinical care and in clinical trials, data analytics must address some of the statistical considerations we have identified through this analysis. In particular, our analysis indicates that symptom severity fluctuates in a manner that might be related to both medication effects and to other factors. This observation raises two important considerations. First, the use of remote assessments to evaluate medication effects—or other effectors relevant to clinical care—must take into account confounding effects that may arise from other factors across a variety of time scales. This can be

addressed with intentional scheduling strategies for data collection and with the application of statistical approaches that carefully consider observed confounders when making inferences. In the presence of unmeasured confounders, one approach that we propose is the use of instrumental variables in randomized experiments with imperfect compliance to detect and estimate personalized treatment effects in mobile health studies³⁸.

Second, the capacity of sensor-based technology to identify changes in assessment performance related to factors other than medicine supports the use of remote digital approaches to support personalized disease management. Because medication was the only contextual annotation collected around assessment performance in mPower, it is not possible to understand other factors causing this ToD effect; these may be biological (circadian rhythms) or technical (routine changes in environment throughout the day). The promise of personalized digital health assessments—and participant-led research—indicates that a broader mechanism needs to be considered that supports annotation and contextualization of such factors to develop actionable recommendations. This may include approaches to collect contextual information through multisensored approaches to explicitly evaluate the relationship between two factors, through passive monitoring measures that classify contextual information, and/or through annotation programs that require participants to report on context.

In addition to confounding, the analysis of digital health studies also needs to account for the longitudinal nature of RWD. In particular, autocorrelation in the repeated measurements can invalidate inferences drawn from standard statistical techniques developed to analyze cross-sectional studies. The potential pitfalls associated with the analysis of longitudinal data have been further illustrated for personalized analyses³⁵ as well as population-level analyses³⁹.

One of the main objectives of the mPower study was to evaluate the utility of open RWD distribution to support development and utilization of effective analytical tools for digital health assessments. In mPower, 67% of participants chose to broadly share their data with qualified researchers (in comparison, 80% out of 216,000 people enrolled in 37 studies using the same consent process chose to share their data broadly). Data collected over the first 6 months of the mPower study was curated and released in conjunction with the beginning of this analysis²⁴. Over 180 individuals from 108 institutions have requested access to the data and dozens of analysis has

been performed that use the mPower data. This community provides an opportunity to evaluate analytical approaches and benchmark appropriate methods for use of sensor-based health data. Informal evaluation across research findings is insightful. When results are comparable across groups, the community is assured as to the validity of results. When results differ across groups, there is a need to further evaluate research approaches. As an example, when comparing our classification results with those of other investigators in the field, we have observed that many have inadvertently not considered confounders (both age and gender, as well as identity confounders).

We have also worked to use the mPower data to develop best practices through a formal benchmarking challenge designed to evaluate methods across the community for the summarization of raw sensor readings into PD severity measures⁴⁰. Because the results described in this analysis were developed by collapsing sensor data into an elementary set of features selected based on previous literature^{41–43}, we anticipated that groups with more sophisticated methods for feature extraction would report improved results. We launched a DREAM challenge that was joined by 400 researchers who competed to derive features for the gait test⁴⁴. The best performing method exhibited a 58% improvement in classification of PD from gait data relative to the results reported in this initial analysis. Successful development of measurements that are demonstrably better than others will require unbiased benchmarking that can compare different methods directly and that can be deployed continually to allow for improvements and generalizability across study designs. As such, the mPower study can provide an open resource that can continue to support benchmarking of emerging methods relative to those that are now best in the field.

The mPower study rapidly achieved enrollment of large numbers of participants but retention in the study declined exponentially (Supplementary Fig. 1). Although there were clear limitations in the study design relating to incentives and communication that likely affected retention, the observed retention rate matches patterns observed across several mobile health applications⁸. Indeed, retention patterns in mobile research studies tend to mirror use patterns for apps in general, rather than retention patterns observed in traditional clinical research studies. This is a principal barrier to the use of technology-based remote research studies. One factor that contributes to this retention issue is that the commitment required to enroll in a remote study is often quite low relative to in-clinic studies. As such, study populations include individuals that vary widely in their commitment to full research participation. This was particularly prominent with mPower and the initial ResearchKit applications that were announced to Apple enthusiasts as part of that company's 2015 promotional event. Reducing the curiosity factor while encouraging long-term engagement could be achieved through technical solutions that, for example, expose study components to nonparticipants to satisfy curiosity or through protocol solutions that, for example, require clinical referrals or other high-touch activities as part of enrollment²². The use of remote studies to serve as registries for enrollment in subsequent studies is also an approach that may address this issue.

Engagement and retention in remote research studies are also affected by an altered value proposition for participation as compared with in-clinic studies. Although risks and benefits for research participation must always be balanced to promote participation, remote studies pose a new set of factors to consider in this value proposition. As an example, studies that require frequently sampled RWD require participants to integrate these study activities into the fabric of their daily lives, competing with other interests and obligations. In addition, many new issues around data generation—including the use of frequently sampled data streams, a participant's own digital data, data collected through commercial devices governed by third-party terms and/or passive monitoring—all contribute to real and perceived changes to the privacy risks associated with

participation. These issues can be mitigated in part by minimizing the breadth or depth of data collection. This can be achieved in many ways, including through study designs that replace continuous monitoring with burst designs and on-device analytics that limit the transfer of high-dimensional data and through mechanisms that allow participants to control data collection. We have taken some of these approaches in a follow-up study called mPower progression, where we have removed overlapping assessments and follow a burst model where 2-week bursts of daily activities are designed to capture the daily fluctuations we observed in this study repeated every 3 months to track long-term progression (<https://parkinsonmpower.org/your-story>). One approach to understand and address the relevant issues that affect participation is to actively incorporate participants into the process of study design^{30,45}. Qualitative content collected from mPower participants²⁴ and through interviews with PD patients indicated that this population was most willing to engage in frequent, remote assessments when that information could be shared back with themselves or their care team to inform disease management^{30,45}. While some research studies may not be able to report results from such activities without biasing research results, this highlights the need to provide value back to research participants in a manner commensurate with the requirements for participation. We also believe that evaluating the effects of engagement interventions through experimentation will be important for the field to evaluate the best methods (<https://parkinsonmpower.org/your-story>).

In summary, the use of smartphones and wearables to collect real-world evidence requires the development of a recalibrated set of research methods as compared to capturing measures in a clinical or laboratory setting. The complexities introduced by remote assessments require careful consideration to support accurate inferences. Despite these limitations, health research conducted through remote studies provides the potential to greatly augment our ability to understand and monitor health. Our results contribute evidence that remote, smartphone-based research studies provide an accessible approach to support frequent, objective and personalized assessments of disease symptoms.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-021-00974-9>.

Received: 10 October 2018; Accepted: 4 June 2021;

Published online: 09 August 2021

References

1. Sherman, R. E. et al. Real-world evidence—what is it and what can it tell us? *N. Engl. J. Med.* **375**, 2293–2297 (2016).
2. Steinhubl, S. R. et al. Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: the mSToPS randomized clinical trial. *JAMA* **320**, 146–155 (2018).
3. Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. Effectiveness of continuous glucose monitoring in a clinical care environment: evidence from the Juvenile Diabetes Research Foundation continuous glucose monitoring (JDRF-CGM) trial. *Diabetes Care* **33**, 17–22 (2010).
4. Anguera, J. A., Jordan, J. T., Castaneda, D., Gazzaley, A. & Areán, P. A. Conducting a fully mobile and randomised clinical trial for depression: access, engagement and expense. *BMJ Innov.* **2**, 14–21 (2016).
5. Quer, G. et al. Home monitoring of blood pressure: short-term changes during serial measurements for 56398 subjects. *IEEE J. Biomed. Health Inform.* **22**, 1691–1698 (2018).
6. Chaibub Neto, E. et al. Learning Disease vs Participant Signatures: a permutation test approach to detect identity confounding in machine learning diagnostic applications. Preprint at <https://arxiv.org/abs/1712.03120> (2017).

7. Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C. & Kording, K. P. The need to approximate the use-case in clinical machine learning. *Gigascience* **6**, 1–9 (2017).
8. Dorsey, E. R. et al. The use of smartphones for health research. *Acad. Med.* **92**, 157–160 (2017).
9. Pratap, A. et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. *NPJ Digit. Med.* **3**, 21 (2020).
10. 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).
11. Espay, A. J. et al. Technology in Parkinson's disease: challenges and opportunities. *Mov. Disord.* **31**, 1272–1282 (2016).
12. Ellis, R. J. et al. A validated smartphone-based assessment of gait and gait variability in Parkinson's disease. *PLoS ONE* **10**, e0141694 (2015).
13. Kostikis, N., Hristu-Varsakelis, D., Arnaoutoglou, M. & Kotsavasiloglou, C. A smartphone-based tool for assessing Parkinsonian hand tremor. *IEEE J. Biomed. Health Inform.* **19**, 1835–1842 (2015).
14. 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).
15. Hauser, R. A. et al. A home diary to assess functional status in patients with Parkinson's disease with motor fluctuations and dyskinesia. *Clin. Neuropharmacol.* **23**, 75–81 (2000).
16. Heldman, D. A. et al. The modified bradykinesia rating scale for Parkinson's disease: reliability and comparison with kinematic measures. *Mov. Disord.* **26**, 1859–1863 (2011).
17. Jeon, H. et al. Automatic classification of tremor severity in Parkinson's disease using a wearable device. *Sensors* **17**, 2067 (2017).
18. Stamate, C. et al. Deep learning Parkinson's from smartphone data. in *Proc. 2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)* 31–40 (IEEE, 2017).
19. Liddle, J. et al. Measuring the lifespan of people with Parkinson's disease using smartphones: proof of principle. *JMIR Mhealth Uhealth* **2**, e13 (2014).
20. Ginis, P. et al. Feasibility and effects of home-based smartphone-delivered automated feedback training for gait in people with Parkinson's disease: a pilot randomized controlled trial. *Parkinsonism Relat. Disord.* **22**, 28–34 (2016).
21. 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).
22. Lipsmeier, F. et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Mov. Disord.* **33**, 1287–1297 (2018).
23. 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).
24. Bot, B. M. et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci. Data* **3**, 160011 (2016).
25. Chaibub Neto, E. et al. On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: the mPower case study. Preprint at <https://arxiv.org/abs/1706.09574> (2017).
26. Chaibub Neto, E., Tummacherla, M., Mangravite, L. & Omberg, L. Causality-based tests to detect the influence of confounders on mobile health diagnostic applications: a comparison with restricted permutations. in *Machine Learning for Health (MLAH) Workshop at NeurIPS 2019*. <https://arxiv.org/abs/1911.05139> (2019).
27. Little, M. A. et al. Using and understanding cross-validation strategies. Perspectives on Saeb et al. *Gigascience* **6**, 1–6 (2017).
28. Neto, E. C. et al. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ Digit. Med.* **2**, 1–6 (2019).
29. Sakar, B. E. et al. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE J. Biomed. Health Inform.* **17**, 828–834 (2013).
30. Mishra, S. R. et al. Supporting coping with Parkinson's disease through self-tracking. in *Proc. 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)* (ACM) 1–16 (2019).
31. Newey, W. K. & West, K. D. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703 (1987).
32. Newey, W. K. & West, K. D. Automatic lag selection in covariance matrix estimation. *Rev. Econ. Stud.* **61**, 631–653 (1994).
33. Grömping, U. Estimators of relative importance in linear regression based on variance decomposition. *Am. Stat.* **61**, 139–147 (2007).
34. Thenganatt, M. A. & Jankovic, J. Parkinson disease subtypes. *JAMA Neurol.* **71**, 499 (2014).
35. Khazaal, Y. et al. Does self-selection affect samples' representativeness in online surveys? An investigation in online video game research. *J. Med. Internet Res.* **16**, e164 (2014).
36. Bethlehem, J. Selection bias in web surveys. *Int. Stat. Rev.* **78**, 161–188 (2010).
37. Badawy, R. et al. Automated quality control for sensor based symptom measurement performed outside the lab. *Sensors* **18**, 1215 (2018).
38. Chaibub Neto, E. et al. Towards personalized causal inference of medication response in mobile health: an instrumental variable approach for randomized trials with imperfect compliance. Preprint at <https://arxiv.org/abs/1604.01055> (2016).
39. Barnett, I., Torous, J., Staples, P., Keshavan, M. & Onnela, J.-P. Beyond smartphones and sensors: choosing appropriate statistical methods for the analysis of longitudinal data. *J. Am. Med. Inform. Assoc.* **25**, 1669–1674 (2018).
40. Marbach, D. et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
41. Tavares, A. L. T. et al. Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation. *Mov. Disord.* **20**, 1286–1298 (2005).
42. Tsanas, A., Little, M. A., McSharry, P. E. & Ramig, L. O. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *J. R. Soc. Interface* **8**, 842–855 (2011).
43. Sejdic, E., Lowry, K. A., Bellanca, J., Redfern, M. S. & Brach, J. S. A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains. *IEEE Trans. Neural Syst. Rehabil. Eng.* **22**, 603–612 (2014).
44. Sieberts, S. K. et al. Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge. *NPJ Digit. Med.* <https://doi.org/10.1038/s41746-021-00414-7> (2021).
45. Doerr, M. et al. Formative evaluation of participant experience with mobile econsent in the app-mediated Parkinson mPower study: a mixed methods study. *JMIR Mhealth Uhealth* **5**, e14 (2017).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Experimental Methods. *Smartphone application.* The study was approved by Western Institutional Review Board (WIRB protocol no. 20141369). The trial registration number is <https://www.clinicaltrials.gov/ct2/show/NCT02696603>. A smartphone application for PD called mPower²⁴ was built on ResearchKit, which is open source and available on GitHub (<https://github.com/Sage-Bionetworks/mPower>). Data from smartphones were transmitted to Sage Bionetworks, a nonprofit organization, via its Bridge Server, and to Synapse²⁴. The smartphone application has three principal components. The first is a set of surveys on the demographics, medical conditions (including whether the individual had been diagnosed with PD), medications for PD (if applicable), and other questions on PD. The second is a series of active assessments, including the following: (1) a speeded tapping test in which participants tap with alternating fingers as fast as they can for 20 s, (2) a voice test in which participants say 'aaah' for 10 s, (3) a gait test in which individuals walk forward for 30 s, (4) a balance test in which individuals stand still for 30 s and (5) a memory test in which individuals recall patterns of alternating lit flowers on a screen. Instructions for each of these assessments are provided as part of the application. The third component is passive monitoring of activity and location using a global positioning system that involves measuring distance from a deidentified location. Because little data from passive monitoring were collected in this study, those results are not presented here. For more details see ref. ²⁴.

Study participants. The study was open to individuals with and without PD. Those who downloaded the mPower application then completed an interactive, in-app informed consent process that included a quiz on the risks, benefits and options for study participation and sharing. Enrollment required correct answers to all questions, and individuals could take the quiz several times. Participants then had to verify their email address to confirm their enrollment⁴⁵.

In-clinic assessments. To enable validation of assessments captured by the mPower application with traditional clinical assessments, we recruited a subset of participants with and without PD to complete up to three in-person assessments (baseline, month 3 and month 6) over 6 months. The recruited individuals performed the assessments on the mPower application in-clinic in addition to completing traditional research assessments, including the MDS-UPDRS⁴⁶, the Montreal Cognitive Assessment⁴⁶, the Timed Up and Go test, Hamilton Anxiety and Depression Scales⁴⁷ and a patient-reported SE-ADL⁴⁸.

In addition, at the 6-month visit, individuals with PD who were taking levodopa were asked to come into clinic in the 'off' state (defined as not having taking their levodopa for at least 12 h) to perform the traditional motor tests and the smartphone assessments and then to take levodopa and perform the same assessments in the 'on' state (at least 30 min after their standard levodopa dose)⁴⁹. The University of Rochester Research Subjects Review Board reviewed and approved this portion of the study.

Feature engineering. The data captured from each of the activities consisted of raw data that was summarized into features. These features were extracted using a feature engineering workflow called mpowertools (<https://github.com/Sage-Bionetworks/mpowertools>) that also has documentation on all of the generated features (<https://github.com/Sage-Bionetworks/mpowertools/blob/master/FeatureDefinitions.md>). Wrapper functions to extract features from data in Synapse database tables using mpowertools are also made available (<https://github.com/Sage-Bionetworks/mPowerRerun/tree/main/R/FeatureExtraction>). Feature extraction for all the activities, except the voice activity, was performed using R programming language v.3.3.3 (ref. ⁵⁰). Voice activity was processed using the Matlab R2016a software. A brief description, by activity type, is provided next.

Tapping. Finger tapping activity measured dexterity, speed and abnormality in kinesis (including hastening, faltering and/or freezing). Participants were instructed to lay their phone on a flat surface and to use two fingers on the same hand to alternatively tap two stationary points on the screen for 20 s. We recorded the screen pixel position (x, y coordinates) of where participants tapped along with time-stamps of when. From the x, y coordinate, a range of features were measured, including the total number of taps measuring the tapping speed, frequency where neither the right or left point was tapped, summary statistics on tapping interval between two points, summary statistics on the drift from each point and correlation between the x, y coordinates as a surrogate for handedness⁵¹.

Voice. Sustained phonation activity measured changes in loudness, pitch, breathiness, roughness and vocal tremor. In this activity participants were instructed to say 'aaah' into the microphone at a steady volume for up to 10 s. The data from this activity include audio files (in m4a format) containing recordings from the iPhone microphone. We extracted a range of summary measures that quantify roughness and aperiodicity, including pitch estimates, change in pitch estimates over time (jitter), change in amplitude over time (shimmer), mel-frequency cepstrum (MFCC) bands and change in MFCC bands over time (MFCC jitter)⁴².

Walk. Walk activity evaluated participants' gait. The first release of the app (v.1.0, build 7) instructed participants to walk 20 steps in a straight line, turn around,

stand still for 30 s then walk 20 steps back. Subsequent releases omitted the return walk. For each leg of this activity, data included measurements from the phones accelerometer, pedometer and gyroscope in both raw and processed formats. Using the three-dimensional (x, y, z) measurements from the accelerometer during the first 30 s, we computed the average acceleration and jerk measurements. Different summary statistics of both the three-dimensional measurements and average acceleration and jerk were used as features for downstream analysis. Furthermore, sophisticated measures, such as the autocorrelation, zero-crossing rate, detrended fluctuation analysis, Teager-Kaiser energy operator, and frequency interval of the Lomb-Scargle periodogram that characterize changes in body motion were also used⁴³.

Balance. Walk activity also had a rest or balance test consisting of standing still. Like walk activity feature sets, summary statistics of average acceleration and detrended fluctuation analysis features were extracted for rest activity. Also, we computed turning time, postural accelerations, postural energy and displacement volume changes characterizing the turning events during the balance activity.

Statistical analyses. *Matching.* Because the PD participants tended to be younger than the non-PD participants, all the population-level analyses were based on a subset of age-matched participants. We performed exact one-to-one age matching of PD and non-PD participants using the MatchIt⁵² R package. Matching was applied separately to male and female participants, and the analyses were based on the pooled set of age-matched females and age-matched males.

Variability comparison between PD and non-PD participants. To compare the variability of PD and non-PD participants, we first calculated the interquartile range (IQR) of each feature of each participant, and then, for each feature, we compared the IQR distribution of PD participants against the IQR distribution of the non-PD participants. (For each participant, the IQR of each feature was computed from all longitudinal measurements provided by the participant.) To remove long-term variation (due to, for instance, learning trends), we detrended the feature data (by subtracting a lowess smoother) before calculating the IQR values. Results were based on a subset of participants that contributed data for at least 15 days (filtered from the set of age/gender-matched participants). Note that, for the participants that performed the activity several times per day, we randomly selected a single record per day to enter the calculation of the IQRs. The number of PD/non-PD participants available for these analysis in the tapping, walk, balance and voice assessments corresponded, respectively, to 67/28, 16/6, 12/5 and 65/23 participants.

For each feature, we assessed if the median IQR of the PD cohort was larger than the median IQR of the non-PD participants using a permutation test, where we shuffled the disease labels and recomputed the median IQR values of the permuted PD and non-PD cohorts. Explicitly, for each feature k , we computed two-tailed permutation P values as, $2 \times \min(P(S_k < s_k), P(S_k > s_k))$, where the test statistic S_k was defined as $S_k = \text{median}_i(\text{IQR}_{ik}) - \text{median}_j(\text{IQR}_{jk})$, where i and j index PD and non-PD participants, respectively, and $P(S_k < s_k)$ and $P(S_k > s_k)$ were estimated as,

$$\frac{1 + \sum_{b=1}^B 1\{S_{kb}^* < s_k\}}{B + 1},$$

and

$$\frac{1 + \sum_{b=1}^B 1\{S_{kb}^* > s_k\}}{B + 1},$$

where B , S_{kb}^* and s_k represent, respectively, the number of permutations, the value of the statistic S_k computed at permutation b , and the observed value of S_k .

Note that, for these analyses, we adopt the IQR statistic because, contrary to other robust measures of dispersion such as the median absolute deviation (MAD), the IQR does not assume that the data is distributed symmetrically. As pointed out in ref. ⁵³, MAD corresponds to finding the symmetric interval (around the median) that contains 50% of the data (which is not a natural way to measure dispersion in asymmetric distributions). The IQR statistic, on the other hand, does not suffer from this issue since the quartiles do not need to be equally far away from the center.

PD versus non-PD classification. For all activity assessments, we built classifiers of PD versus non-PD participants using both the random forest algorithm⁵⁴ implemented in the randomForest⁵⁵ R package (with the default tuning parameter settings) and the ridge-regression classification⁵⁶ implemented in the glmnet R package⁵⁷ (using fivefold cross-validation optimization implemented in the cv.glmnet function). For the ridge-regression classification, we also standardized the features (by subtracting the mean and dividing by the s.d.) before running the analyses. Classification performance was evaluated using the AUC, and (balanced) accuracy metrics. The classifiers were trained and evaluated on 100 distinct random splits of the data into training and test sets, with half of the data used for training and half for testing.

Due to imbalance in demographic characteristics of the PD and non-PD participant cohorts, as well as to the availability of longitudinal data contributed by the participants, we need to be cautious about the potential sources of confounding, which we describe next.

Evaluating education, gender and age confounding. In statistics and epidemiology, a confounder is defined as a variable that influences both the 'dependent variable' (or 'exposure' or 'treatment') and the 'response' (or 'independent') variable, generating an spurious association between the dependent and response variables. Confounding is a causal concept and, therefore, cannot be described solely in terms of associations. Its definition requires the use of causal diagrams describing our qualitative assumptions about the causal relations between the variables⁵⁸. (See also Chapter 4 of ref. ⁵⁹ for a gentle introduction to the topic.)

Following the methodology proposed in ref. ²⁶ (described in Supplementary Note 1), we used causality-based confounding tests to evaluate if the predictive performance of a classifier is influenced by measured confounders. The basic idea is to consider the causal diagram describing the data generation process behind the classification task (which is known in our application) and evaluate if the CI relationships observed in the data agree with the CI relations implied by the causal diagram. The goal is to evaluate if demographic variables that are associated with the disease labels (such as age, gender and education in the mPower data) are indeed confounding the predictions of classifiers of disease status (PD versus non-PD). To this end, we train classifiers (using unmatched data) and evaluate the CI relationships between the confounders, the labels and the classifier predictions (that is, the predicted probability that a test set example is a PD case) in the test set. Only the variables that are indeed confounding the predictions need to be used for the matching adjustment. Supplementary Note 1 provides a detailed description of the approach, as well as its application to the education, gender and age variables in the mPower data. The methodology and results are presented in Supplementary Figs 11–30. We use standard two-tailed (partial) correlation tests (based on Pearson correlations) and one-tailed (partial) distance correlation tests^{60,61} for testing for CI relationships. (Note that the partial distance correlation tests are one-tailed because distance correlation assumes values only between 0 and 1.)

Detecting identity confounding. Recent work in diagnostic machine learning^{25,62,63}, showed that the accuracy of a classifier can be overestimated when the training/test sets are generated using a record-wise data split strategy (that is, when a randomly chosen subset of the records contributed by a participant is assigned to the training set, while the remaining records end up in the test set). The reason is that having data of the same participant in both the training and test sets encourages the classifier to recognize subject-specific patterns in the test set that the algorithm learned from the training set, in addition to (or instead of) learning about the disease characteristics. In other words, the feature-to-disease relationship learned by the classifier is confounded by the identity of the subjects, and the classifier might end up performing subject identification rather than disease recognition.

We assessed the presence of identity confounding in our datasets using subject-wise label permutations (where all the records of a given participant are either assigned the PD or the non-PD label, before splitting the data into training and test sets, during the permutation process). Explicitly, for each activity assessment, our analyses were based on the following two-step strategy:

- First, we generated a disease recognition null distribution²⁵ by randomly shuffling the participant's labels (PD versus non-PD) in a subject-wise manner 1,000 times, and calculating the AUC in each permuted dataset (using the same record-wise data split across all 1,000 label shufflings). The presence of identity confounding was assessed informally by inspecting if this null distribution was centered at an AUC value larger than 0.5 (the baseline random guess value for the AUC metric). Note that identity confounding leads to shift from the 0.5 baseline AUC value because the subject-wise label shuffling strategy breaks the association between labels and feature data, but still preserves the association between features and participant identities, so that it neutralizes the classifier ability to recognize disease, but still allows the classifier to learn about the participant's characteristics. The blue histograms in Supplementary Fig. 2a–d show the disease null distributions for the tapping, walk, rest and voice classifiers, respectively.
- Second, we computed the pseudo *P* value statistic proposed in ref. ²⁵, which corresponds to the probability of observing an AUC value larger than the median of the disease recognition null distribution under the null hypothesis that the classifier is not performing disease recognition and subject identification. The pseudo *P* values for the tapping, walk, rest and voice classifiers are depicted, respectively, as the area under the black density to the right of the black line in Supplementary Fig. 2a–d. A small pseudo *P* value indicate strong evidence against the null hypothesis of no identity confounding.

To assess the robustness of the results to the training/test data splits, we report the results from 30 distinct random splits of the data in Supplementary Fig. 2e–h.

Classification performance using collapsed features and subject-wise train/test data splits. Due to the presence of strong age confounding (and slight gender confounding) we adopted an age-/gender-matched subset of participants

(described above) for the PD versus non-PD classification analyses. Furthermore, due to the presence of strong identity confounding in the data, we evaluated the classification performance in two distinct ways: by collapsing the longitudinal feature data of each participant into median and IQR summary values as well as by adopting a subject-wise train/test data split, in which all longitudinal data (all records) of each participant were assigned to either the training or to the test set. Note that both strategies avoid the identity confounding issue described above. Furthermore, the collapsed feature strategy seems to have the extra advantage of improving generalization, as reported before in the literature⁶⁴ (and also observed by us in our analyses). To assess the robustness of the results to the training/test data splits, we report the results from 100 random splits of the data (Supplementary Figs. 3 and 4). Furthermore, to assess if the classification performance is better than a random guess, we also present results based on shuffled labels, which provide information about the range of AUC and balanced accuracy scores we can expect to see by chance.

For the classifiers based on collapsed features, we also provide analytical *P* values for testing if the AUC score is >0.5. Explicitly, it has been shown⁶⁵ that, when there are no ties in the predicted class probabilities used for the computation of the AUC, the test statistic of the Mann-Whitney *U* test (*U*) is related to the AUC statistic by, $U = n_n n_p (1 - \text{AUC})$, where n_n and n_p represent the number of negative and positive labels in the test set (see section 2 of ref. ⁶⁶ for details).

In the presence of ties, the *P* value can be computed as the left tail of the asymptotic approximate null,

$$U \approx N \left(\frac{n_n n_p}{2}, \frac{n_n n_p (n + 1)}{12} - \frac{n_n n_p}{12 n (n - 1)} \sum_{j=1}^{\tau} t_j (t_j - 1) (t_j + 1) \right),$$

where $n = n_n + n_p$, τ is the number of groups of ties, and t_j is the number of ties in group j ⁶⁶.

Alternatively, we can get the *P* value as the right tail probability of the corresponding AUC null, $\text{AUC} \approx N(0.5, \phi^2)$,

$$P\text{value} = 1 - \Phi \left(\frac{\text{AUC}_0 - 0.5}{\phi} \right),$$

where $\Phi()$ represents the cumulative distribution function of a standard normal distribution, and

$$\phi^2 = \frac{n + 1}{12 n_n n_p} - \frac{1}{12 n_n n_p n (n - 1)} \sum_{j=1}^{\tau} t_j (t_j - 1) (t_j + 1).$$

Analyzing personalized medication response and daily fluctuations. Data processing.

In our personalized analyses, we investigate the influence of both medication and time-of-the-day (ToD) effects. Because ToD is a circular variable, we have that the linear term used in our models for encoding this variable will treat values such as 11:59 p.m. and 00:01 a.m. very differently (even though these values are only 2 min apart). To avoid potential issues arising from the circularity of the ToD variable, we filtered out any activities (records) that were performed between midnight and 5 a.m.

Every time a participant performed an activity assessment, the participant was asked if the activity was done 'Immediately before Parkinson medication', 'Just after Parkinson medication (at your best)', 'Another time' or 'I don't take Parkinson medications'. For each of the activity assessments we used this self-reported information to select participants that performed at least 15 assessments before taking medication (that is, 'Immediately before Parkinson medication') and at least 15 assessments after taking medication (that is, 'Just after Parkinson medication (at your best)'). Supplementary Fig. 7 reports the number of assessments per participant (both before and after medication) used in the analyses.

For each participant, the data from each extracted feature was separately detrended with a lowess smoother (so that our feature data, actually corresponds to the residuals of a lowess fit to the data point collection index). (Detrending was necessary to avoid artifacts, such as detecting a difference between activities performed before and after medication due to the learning trend, in participants that tended to perform activities before medication at a higher frequency in the beginning of the study, before switching to performing after medication activities at a higher frequency later on, or vice versa.) The data were also transformed to a roughly normal distribution using a rank-quantile transformation,

$$\Phi^{-1} \left(\frac{r_i - 0.5}{n} \right),$$

where $\Phi()$ represents the cumulative density function of the standard normal random variable, r_i represents the rank of the outcome value, y_i , and n represents the number of outcome data points.

Detecting putative treatment and ToD effects. One of our goals was to determine whether a particular patient is responding to dopaminergic medication. However, because mPower is an observational study, any association between

treatment (medicated versus unmedicated) and outcome (performance on the activity assessment) might be due to unmeasured confounders, rather than to a causal effect of the treatment on the outcome. In particular, causal inferences at the personalized level are especially vulnerable to confounding effects that arise in a cyclic manner across the day (such as circadian rhythms and daily routine activities). For instance, our data shows that some participants tended to perform the 'before medication' activity assessments earlier in the day than the 'after medication' assessments (Supplementary Fig. 31). For these participants, we cannot directly conclude that an observed improvement in performance between assessments performed before versus after medication are indicative of a medication effect, as the difference in performance might be due to daily cyclic confounders.

A special characteristic of mobile health studies is that the ToD that the activity is performed is always recorded by the smartphone. Because ToD can be potentially used as a surrogate measurement for circadian rhythms and daily routine confounders, it is possible to disentangle the putative effects of medication and cyclic daily confounders on the outcome. As fully described in ref. ²⁵, it is possible to determine if a participant is likely responding to the medication or to the ToD, or both, using the strategy summarized next.

Following the methodology presented in ref. ²⁵, for each separate feature we first disentangle putative treatment effects from putative ToD effects by considering distinct equivalence classes of causal models (involving the treatment, ToD and outcome variables), and selecting the class that is better supported by the data, using CI tests implemented via *t*-tests in linear regression models for time series data. Next, we combine the results across all features into a union-intersection test, where we test a global null hypothesis of no putative treatment effect for any of the extracted features versus the alternative hypothesis that there is a putative treatment effect for at least one of the features. (And, similarly, for the putative ToD effect.) See ref. ²⁵ for details. One important assumption of this approach is that there are no unmeasured confounders influencing the treatment, ToD and the outcome variables. (This is why we qualify the effects as 'putative treatment' and 'putative ToD' effects. While ToD is the most important confounder of the treatment/outcome relationship in personalized analyses, it might still be possible that the observed associations between these variables might be generated to some extent by other unobserved confounders.)

To account for potential serial associations in the data, we repeated our analyses using three distinct regression-based approaches (which account for residual autocorrelation in different ways) including: (1) regression modeling with heteroscedasticity, and autocorrelation consistent (HAC) covariance matrix estimation, based on the Newey–West HAC estimator⁶⁷ (using Bartlett kernel, and the automatic bandwidth selection procedure described in ref. ³², and implemented in the sandwich R package⁶⁸); (2) regression modeling with autoregressive integrated moving average⁶⁹ (ARIMA) errors (using the auto.arima function of the forecast R package⁷⁰ to first select the autoregressive, moving average and differencing orders of the models); and (3) a standard linear regression approach with independent and identically distributed Gaussian errors (no serial association adjustment).

While both ARIMA and Newey–West HAC estimation assume the data is equally spaced (which is not true in our application), studies have shown that application of the Newey–West estimator to time series with missing data (and, hence, unequally spaced) still generates asymptotically consistent estimates of the covariance matrix, as well as reasonable performance in finite sample simulation studies^{71,72}. For this reason, we report the Newey–West HAC estimator results in the main text. Results based on the other approaches are largely consistent, although the ARIMA and Newey–West approaches tended to detect a larger number of significant findings than the standard linear regression. (Note that this is in accordance with theoretical results⁷³ that show that when there is positive autocorrelation in paired data, the *F*-tests (*t*-tests) tend to be conservative. See ref. ²⁵ for further details and illustrations.)

The results presented in Fig. 2 and Supplementary Fig. 8 were based on two rounds of multiple testing correction. The first is performed across all features, during the computation of the union-intersection *P* value of each individual (as described in ref. ²⁵). The second is performed across the individuals, by applying multiple testing correction to the union-intersection *P* values. In both rounds of correction, we adopted the Benjamini–Hochberg multiple testing correction procedure⁷⁴.

Furthermore, to provide a measure of effect sizes, we also performed a relative importance analysis³³ of our personalized models. The basic idea is to decompose the *R*² statistic, which measures how much of the variability of the response (feature) is explained by the covariates (medication and ToD), into the separate contributions of the medication and ToD covariates. (Note that the relative importance contributions of each covariate sum up to the *R*² value.) Following recommendations in ref. ³³, we adopted the method proposed by refs. ^{75,76}, which estimates the relative importance of the variables using sequential sum of squares averaged across all possible orderings of the covariates. (In our application, there are only two possible orderings of the covariates, namely, feature ~ medication + ToD and feature ~ ToD + medication. The ~ symbol is notation for 'regressed on'.) In our analyses, we used the implementation provided in the relaimpo R package⁷⁷ (based on the lmg option).

Figure 2 reports the distribution of the relative importance for medication and of ToD faceted by fluctuators versus nonfluctuator participants. Note that, because the personalized analyses reported in Fig. 2 are based on a union-intersection test (where, for each participant and activity, the union-intersection *P* value corresponds to the minimum *P* value across all features), we report the output of the relative importance analyses for the corresponding most significant feature in each activity. For instance, for the medication analysis reported in Fig. 2c, if the most significant feature for the first participant in the tapping activity is, for example, the number of taps, then we report the relative importance of the medication covariate calculated from the linear model that uses the number of taps as the response variable. Similarly, for the ToD analysis reported in Fig. 2f, if the most significant feature of the first participant is, for example, the mean tapping interval, then we report the relative importance of the ToD covariate from the linear model that uses mean tapping interval as the response.

Consistency between differences in PD and non-PD participants and the direction of medication effects. To verify whether the direction of the putative medication effects were consistent with the difference observed between PD and non-PD participants, for each of the 186 features across the four activity tasks, we computed the differences,

$$\Delta PD(\text{feature}) = \text{median}_{PD}(\text{feature}) - \text{median}_{nonPD}(\text{feature}),$$

where $\text{median}_{PD}(\text{feature})$ represents the median of the feature values computed among the PD participants, and $\text{median}_{nonPD}(\text{feature})$ represents the respective quantity computed among the non-PD participants, and compared the sign of ΔPD against the sign of the averaged medication effect (computed across the PD participants that entered our personalized analyses). Note that, for the computation of ΔPD , for each feature, we first collapsed the longitudinal data of each participant into its median value and then used these collapsed median values in the computation of $\text{median}_{PD}(\text{feature})$ and $\text{median}_{nonPD}(\text{feature})$.

It is important to point out that consistent ΔPD s and medication effects have opposite signs. To see why, note that for features where higher values indicate better performance (such as number of taps), a negative ΔPD indicates that PD participants tend to do worse than non-PD participants. For the putative medication effects on the other hand, better performance on these features is captured by positive effects (since our regression models use 'before medication' as the baseline level for the medication variable, so that positive effects indicate higher values after taking medication). As a consequence, for any feature where higher values indicate better performance we have that negative ΔPD s will be consistent with positive medication effects. For example, PD participants tend to show lower number of taps in comparison with non-PD participants, generating, therefore, a negative ΔPD sign. On the other hand, PD participants who respond well to the medication tend to show higher number of taps after taking medication, leading to positive signs for the putative medication effects.

Similarly, for features where higher values indicate worse performance (for example, length of the interval between taps) we have that positive signs for ΔPD will be consistent with negative putative medication effects. For example, PD participants tend to show longer intervals between taps in comparison with non-PD participants, generating, therefore, a positive ΔPD sign. On the other hand, PD participants who improve with medication tend to show shorter intervals between taps after taking medication, leading to negative signs for the putative medication effects.

Hence, for this analysis, we measure consistency in the direction of the putative medication effects and ΔPD , by counting the number of opposite signs between these quantities across all 186 features.

Association between in-clinic measures and the severity score derived from mPower activities. We compared the mPower symptom severity score with three in-clinic PD severity measures, namely, the total MDS-UPDRS, the SE-ADL and the Hoehn and Yahr score. The mPower symptom severity score was derived as the average of the prediction probabilities from each of the classifiers for the PD versus non-PD (one for each of the smartphone assessments, tapping, walk, balance and voice), where the classifier was trained in the mPower data and the predictions were generated for the ObjectivePD cohort data. The trained classifiers and prediction probabilities were generated using collapsed features, where we summarized repeated measures for each participant into the median and IQR values. Association tests were done using Pearson correlation and an independent samples two-tailed *t*-test for the Hoehn and Yahr stage, both using Bonferroni multiple testing correction.

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

Data availability

The raw data from 67% of participants who have chosen to share their data broadly with all qualified researchers is available at <https://doi.org/10.7303/syn4993293>. Features, intermediate results and trained models for these participants are also available in Synapse (<https://doi.org/10.7303/syn23277418>). Access to data requires users to have their Synapse account validated, submit a data use

statement and agree to terms of use. To aid in reproducibility and provide all final and intermediate results to the research community, we have redone the analysis presented here using the broadly shared data.

Code availability

The code, a docker container with all installed packages and a snakemake script that reproduces all of the figure and analysis, is available in GitHub: <https://github.com/Sage-Bionetworks/mPowerRerun>.

References

46. Gill, D. J., Freshman, A., Blender, J. A. & Ravina, B. The Montreal cognitive assessment as a screening tool for cognitive impairment in Parkinson's disease. *Mov. Disord.* **23**, 1043–1046 (2008).
47. Weintraub, D., Oehlberg, K. A., Katz, I. R. & Stern, M. B. Test characteristics of the 15-item geriatric depression scale and Hamilton depression rating scale in Parkinson disease. *Am. J. Geriatr. Psychiatry* **14**, 169–175 (2006).
48. Schwab, R.S. & England, A.C. Projection technique for evaluating surgery in Parkinson's disease. in *Third Symposium on Parkinson's Disease* (eds Gillingham, F. J. & Danoldson, I. M. L.) 152–157 (E & S. Livingston, 1969).
49. McRae, C., Diem, G., Vo, A., O'Brien, C. & Seeberger, L. Schwab & England: standardization of administration. *Mov. Disord.* **15**, 335–336 (2000).
50. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/> (2014).
51. Taylor, A. L. T. et al. Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation. *Mov. Disord.* **20**, 1286–1298 (2005).
52. Ho, D., Imai, K., King, G. & Stuart, E. A. MatchIt: non-parametric preprocessing for parametric causal inference. *J. Stat. Softw.* **42**, 1–28 (2011).
53. Rousseeuw, P. J. & Croux, C. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.* **88**, 1273–1283 (1993).
54. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
55. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
56. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2001).
57. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
58. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge University Press, 2009).
59. Pearl, J. & Mackenzie, D. *The Book of Why: the New Science of Cause and Effect* (Basic Books, 2018).
60. Szekely, G., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **35**, 2769–2794 (2007).
61. Szekely, G. & Rizzo, M. L. Partial distance correlation with methods for dissimilarities. *Ann. Stat.* **42**, 2382–2412 (2014).
62. Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C. & Kording, K. P. The need to approximate the use-case in clinical machine learning. *GigaScience* **6**, 1–9 (2017).
63. Chaibub Neto, E. et al. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ Digit. Med.* **2**, 99 (2019).
64. Sarkar, B. E. et al. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE J. Biomed. Health Inform.* **17**, 828–834 (2013).
65. Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* **12**, 387415 (1975).
66. Mason, S. L. & Graham, N. E. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. *Q. J. R. Meteorol. Soc.* **128**, 2145–2166 (2002).
67. Newey, W. K. & West, K. D. A simple, positive-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703708 (1987).
68. Zeileis, A. Econometric computing with HC and HAC covariance matrix estimation. *J. Stat. Softw.* **10**, 1–17 (2004).
69. Box, G., Jenkins, G. M. & Reinsel, G. C. *Time Series Analysis: Forecasting and Control* 3rd edn (Prentice-Hall, 1994).
70. Hyndman, R. J. & Khandakar, Y. Automatic time series forecasting: the forecast package for R. *J. Stat. Softw.* **26**, 1–22 (2008).
71. Datta, D. D. & Du W. Nonparametric HAC estimation for time series data with missing observations. International Finance Discussion Papers. The Federal Reserve Board (2012).
72. Rho, S. H. & Vogelsang, T. J. Heteroskedasticity autocorrelation robust inference in time series regressions with missing data. *Econometric Theory* **35**, 601–629 (2019).
73. McGregor, J. R. & Babb, J. C. Serially correlated differences in the paired comparison of time series. *Biometrika* **76**, 735–739 (1989).
74. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
75. Lindeman, R. H., Merenda, P. F., and Gold, R. Z. *Introduction to Bivariate and Multivariate Analysis* (Scott, Foresman, 1980).
76. Kruskal, W. Relative importance by averaging over orderings. *Am. Stat.* **41**, 6–10 (1987).
77. Gromping, U. Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw.* **17**, 1–27 (2007).

Acknowledgements

This work was funded through a grant from the Robert Wood Johnson Foundation. These data were contributed by users of the Parkinson mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse (<https://doi.org/10.7303/syn4993293>).

Author contributions

L.O., E.C.N. and L.M.M. wrote the paper. L.O. oversaw the analytical and feature extraction activities. L.M.M. was the principal investigator on the study. E.C.N. developed analytical methods. E.C.N. performed the analyses of the mPower data, assisted by T.M.P., A.P. and L.O. A.T. independently reproduced analyses of the mPower data. T.M.P., A.P. and E.C.N. developed features extraction pipelines and figures. B.M.B. curated data and consulted on analyses. A.K. helped design in-app language, look and logic. M.R.K. led the team that designed and implemented the mPower app, and contributed to study design and data capture methodology. E.R.D. designed and oversaw the ObjectivePD validation study with assistance from M.E. and R.S. C.S. and J.W. led the team that designed and implemented the governance for informed consent and data sharing. A.D.T. conceived the study and helped design the app. J.A. performed in-clinic data collection. B.R.B., S.M.G., K.K., M.A.L., C.T. and C.M.T. served as scientific advisors on the study. All authors assisted with revisions of the paper.

Competing interests

B.R.B. currently serves as Editor-in-Chief for the *Journal of Parkinson's Disease*, serves on the editorial boards of *Practical Neurology* and *Digital Biomarkers*, has received honoraria from serving on the scientific advisory board for Zambon, Biogen, UCB and Walk with Path, has received fees for speaking at conferences from AbbVie, Zambon, Roche, GE Healthcare and Bial, and has received research support from the Netherlands Organization for Scientific Research, the Michael J. Fox Foundation, UCB, Abbvie, Zambon, the Stichting Parkinson Fonds, the Hersenstichting Nederland, the Parkinson's Foundation, Verily Life Sciences, Horizon 2020, the Topsector Life Sciences and Health and the Parkinson Vereniging. E.R.D. has received honoraria for speaking at American Academy of Neurology courses, American Neurological Association and University of Michigan; received compensation for consulting services from 23andMe, Abbott, Abbvie, American Well, Biogen, BrainNeuroBio, Clintrex, Curasen Therapeutics, DeciBio, Denali Therapeutics, GlaxoSmithKline, Grand Rounds, Karger, Lundbeck, MC10, MedAvante, Medical-legal services, Mednick Associates, National Institute of Neurological Disorders and Stroke, Olson Research Group, Optio, Origen Data Sciences, Inc., Otsuka, Prilenia, Putnam Associates, Roche, Sanofi, Shire, Spark, Sunovion Pharma, Teva, Theravance, UCB and Voyager Therapeutics; research support from Abbvie, Acadia Pharmaceuticals, AMC Health, Biosensics, Burroughs Wellcome Fund, Davis Phinney Foundation, Duke University, Food and Drug Administration, GlaxoSmithKline, Greater Rochester Health Foundation, Huntington Study Group, Michael J. Fox Foundation, the NIH/NINDS, NSF, Nuredis Pharmaceuticals, Patient-Centered Outcomes Research Institute, Pfizer, Prana Biotechnology, Raptor Pharmaceuticals, Roche, Safra Foundation, Teva Pharmaceuticals and University of California Irvine; editorial services for Karger Publications; and ownership interests with Grand Rounds (second opinion service). C.M.T. reports grants from Sage Biometrics, during the conducting of the study; grants from Parkinson Foundation, grants from Gateway LLC, grants from Roche/Genentech, grants from Parkinson Study Group, personal fees from Biogen Idec, personal fees from Acorda, personal fees from Adamas Therapeutics, personal fees from Amneal, personal fees from CNS Ratings, personal fees from Grey Matter LLC, grants from Michael J. Fox Foundation, grants from NIH/NIA, grants from NIH/NINDS, grants from VA Merit, grants from Department of Defense, personal fees from Northwestern University, personal fees from Partners, Harvard University, nonfinancial support from Medtronic, Inc., nonfinancial support from Acadia, nonfinancial support from Boston Scientific, nonfinancial support from Neurocrine, nonfinancial support from Acadia, grants from Biogen Idec research support, nonfinancial support from Biogen Idec, personal fees from Guidemark Health, personal fees from Acadia, personal fees from Neurocrine, personal fees from Lundbeck, personal fees from Cadent, nonfinancial support from Neurocrine and grants from Roche Genentech, outside the submitted work.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-021-00974-9>.

Correspondence and requests for materials should be addressed to L.O., E.C.N. or L.M.M.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Data was collected using the purpose built mPower application. mPower source code is available at <https://github.com/Sage-Bionetworks/mPowerSDK>

Data analysis Data was analyzed using R version 3.3.1 with some feature extraction using Matlab R2016a. The code, a docker container with all installed packages and a snakemake script that reproduces all of the figure and analysis is available in GitHub: <https://github.com/Sage-Bionetworks/mPowerRerun>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw data from 67% of participants who have chosen to share their data broadly with all qualified researchers is available at doi:10.7303/syn4993293. Features, intermediate results and trained models for these participants are also available in Synapse [10.7303/syn23277418]. Access to data requires users to have their Synapse account validated, submit a data use statement and agree to terms of use. To aid in reproducibility and provide all final and intermediate results to the research

community we have redone the analysis presented here using the broadly shared data. The code, a docker container with all installed packages and a snakemake script that reproduces the input data for and all figures and analysis is available in GitHub.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In total, the study enrolled 1,414 participants with self reported Parkinson's Disease and 8,432 non-PD participants. Participants self-selected to participate in the study. Therefore, we had no control over sample size prior to collecting the data. 960 participants enrolled and performed at least 5 self-administered active assessments designed to evaluate PD symptoms.
Data exclusions	Individuals who performed less than 5 repeat measures were excluded from analysis as we needed repeat measures to estimate variance in performance. Certain analysis was performed on a gender matched sub-cohort (see methods for details).
Replication	In order to avoid coding mistakes, all analysis were performed independently by at least two individuals using their own computer code. Replication was successful.
Randomization	This was a observational study with open enrollment as such no randomization in the cohort was performed. Analysis was dependent on random selection, bootstrapping and permutations however.
Blinding	This was an observational study without interventions as such no blinding was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	This study consisted of two populations, a larger cohort of participants with and without Parkinson's disease who joined by downloading the study application and a smaller cohort (N=44) who were recruited to perform activities in a clinic as well as the smartphone activities. The population characteristics such as age, gender, education, disease state and medication intake were evaluated as confounders in the analysis. Summaries of distribution of these variables is included in Table 1 of the paper.
Recruitment	Participants self-selected to participate in this observational study. As is usually the case in digital health studies run an uncontrolled environment, selection bias is a major issue. For instance, the control participants tended to be younger than the cases. We employed (i) statistical method to detect and quantify the influence of confounders; and (ii) in order to remove confounding biases we employed matching techniques to reduce the association between potential confounders and disease labels prior to performing the population level analysis.
Ethics oversight	The study was approved by Western Institutional Review Board (WIRB protocol #20141369). The sub-study that included in clinic assessments was approved by The University of Rochester Research Subjects Review Board. All completed an electronic informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.