RESEARCH ARTICLE

Design and results of a smartphone-based digital phenotyping study to quantify ALS progression

James D. Berry[1,2], Sabrina Paganoni[1,2,3], Kenzie Carlson[4], Katherine Burke[2,5], Harli Weber[2], Patrick Staples[4], Joel Salinas[1,2,6], James Chan[7], Jordan R. Green[5], Kathryn Connaghan[5], Josh Barback[4] & Jukka Pekka Onnela[4]

[1]School of Medicine, Harvard Medical School, Boston, Massachusetts
[2]Neurological Clinical Research Institute, Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts
[3]Department of Physical Medicine and Rehabilitation, Spaulding Rehabilitation Hospital, Boston, Massachusetts
[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts
[5]MGH Institute of Health Professions, Charlestown, Massachusetts
[6]Department of Neurology, Henry and Allison McCance Center for Brain Health, Massachusetts General Hospital, Boston, Massachusetts
[7]Department of Biostatistics, Massachusetts General Hospital, Boston, Massachusetts

Correspondence

James D. Berry, Neurological Clinical Research Institute (NCRI), Massachusetts General Hospital, 165 Cambridge St, 6th Floor, Boston, MA 02114. Tel: +1 617 726 5097; Fax: +1 (617) 724-7290; E-mail: jdberry@mgh.harvard.edu

Abstract

Objective: The amyotrophic lateral sclerosis (ALS) trial outcome measures are clinic based. Active and passive smartphone data can provide important longitudinal information about ALS progression outside the clinic. **Methods**: We used Beiwe, a research platform for smartphone-based digital phenotyping, to collect active (self-report ALSFRS-R surveys and speech recordings) and passive (phone sensors and logs) data from patients with ALS for approximately 24 weeks. In clinics, at baseline and every 3 months, we collected vital capacity, ALSFRS-R, and ALS-CBS at enrollment, week 12, and week 24. We also collected ALSFRS-R by telephone at week 6. **Results**: Baseline in-clinic ALSFRS-R and smartphone self-report correlation was 0.93 ($P < 0.001$). ALSFRS-R slopes were equivalent and within-subject standard deviation was smaller for smartphone-based self-report (0.26 vs. 0.56). Use of Beiwe afforded weekly collection of speech samples amenable to a variety of analyses, and we found mean pause time to increase by 0.02 sec per month across the sample. **Interpretation**: Smartphone-based digital phenotyping in people with ALS is feasible and informative. Self-administered smartphone ALSFRS-R scores correlate highly with clinic-based ALSFRS-R scores, have low variability, and could be used in clinical trials. More research is required to fully analyze speech recordings and passive data, and to identify optimal digital markers for use in future ALS clinical trials.

Introduction

Amyotrophic lateral sclerosis (ALS) is neurodegenerative motor neuron disease leading to progressive muscle weakness, dysarthria, dysphagia, and respiratory insufficiency. There is an urgent need to develop effective therapies. Although ALS trials are proliferating, methods are critically needed to maximize participant retention and participation. ALS is a rare and fatal disease, and the burden of frequent clinic visits can lead to high participant attrition in therapeutic trials. Smartphone-based data collection is one promising solution for improving participation in ALS trials. Smartphones are now nearly ubiquitous in North America and provide a latent source of rich data about human behavior. If demonstrated to be efficacious (e.g., valid, reliable), smartphone-based outcome measures could significantly accelerate ALS clinical trials by increasing statistical power and reducing data missingness, trial costs, and participant burden in ALS trials.

Digital phenotyping is defined as the "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices, in particular smartphones".[1,2] It can rely on actively acquired data from tasks (e.g., surveys, finger

tapping, and speech recordings) and passively acquired data from sensors (e.g., GPS and accelerometer) and logs (e.g., phone-use logs and communication logs) without any impact on the daily routines of participants. Active and passive data are complementary and can be validated against more traditional clinic-based outcome measures.

We developed the Beiwe research platform for smartphone-based digital phenotyping to facilitate the discovery and analysis of digital social, behavioral, and cognitive markers. The platform's front-end consists of smartphone applications for Android and iOS devices for data collection; the back-end makes use of Amazon Web Services cloud computing, performs data collection, storage and analysis, and provides study management and monitoring tools. Beiwe records raw (unprocessed) data from smartphone sensors and logs. This provides maximal data flexibility and transparency, and enables pooling of data across studies. The data analysis pipeline, part of the Beiwe back-end, includes algorithms developed specifically to process raw data captured by Beiwe. They provide meaningful information about sleep, social interactions, mobility, gross motor activity, cognitive functioning, and speech production. To facilitate validation and replication studies, Beiwe captures all study settings in human-readable JavaScript Object Notation (JSON)-formatted configuration files, which allows study replication with the exact settings from prior studies. To encourage collaboration and reproducibility, the Beiwe platform is available as an open source software.[3]

We conducted a pilot digital phenotyping study in people with ALS to (1) determine the correlation of smartphone-based (self-report) revised ALS Functional Rating Scale (ALSFRS-R) with standard clinic-based ALSFRS-R administered to patients by staff, (2) quantify changes in behaviors over time as measured by passively collected smartphone data, and (3) determine the feasibility of acquiring speech audio recordings at weekly intervals to evaluate speech degradation over time.

This manuscript focuses on two types of active smartphone data: surveys and audio recordings. Future analyses will focus on more detailed analysis of passive data and speech audio recordings.

Methods

This single-center pilot study was approved by the Institutional Review Board (IRB) of Partners Healthcare prior to initiation. Data analysis for the Beiwe platform was approved by the Harvard University IRB. Participants provided written informed consent prior to initiation of any study procedures or data collection. All data were collected and stored in compliance with Partners and Harvard policies and regulations as well as state and national laws and regulations. Data collection, management, and security were reviewed by the Massachusetts General Hospital (MGH) Information Security Office prior to study initiation.

Data collection, storage and security

Clinical data were stored in an electronic data capture system using standardized clinical data collection, secure storage, and active data monitoring to facilitate data analysis and sharing.

Study personnel assigned each participant a randomly generated Beiwe User ID consisting of eight alphanumeric characters and a temporary password, and they assisted participants with app installation and activation at the time of enrollment.

Data collected by the Beiwe platform are immediately encrypted and stored on the smartphone until the phone is connected to Wi-Fi to avoid use of participants' cellular data plan. Passive data sources were hashed to protect patient identity. All data were encrypted while stored on the phone awaiting upload, while in transit, and while on the server. The data were accessible through the Beiwe platform only to authorized study investigators.

Patient recruitment

Participants were recruited from the ALS Multidisciplinary Clinic at MGH. Participation required a confirmed diagnosis of ALS meeting El-Escorial Criteria, and the ability to provide informed consent. All participants were over 18 years of age, owned a smartphone running an iOS or Android operating system, reported at least moderate smartphone use, and were able to comply with study procedures per the site investigator's assessment.

Study overview

The study diagram is shown in Figure 1. Throughout the study, two forms of clinical data were collected: (1) clinic-based data were obtained by study staff either in-person or via a telephone call, (2) smartphone-based data were obtained using the Beiwe smartphone application.

Following the informed consent process, study personnel collected baseline clinical information from the participants, including medical history, detailed history of the onset and diagnosis of ALS, and baseline clinic-based outcome measures. Study staff also provided participants with standard training in the use of the Beiwe app and the methods for filling out surveys and making audio recordings. From the point of installing the application on their personal phones at the clinic, Beiwe started collecting smartphone-based, self-entered ALSFRS-R item
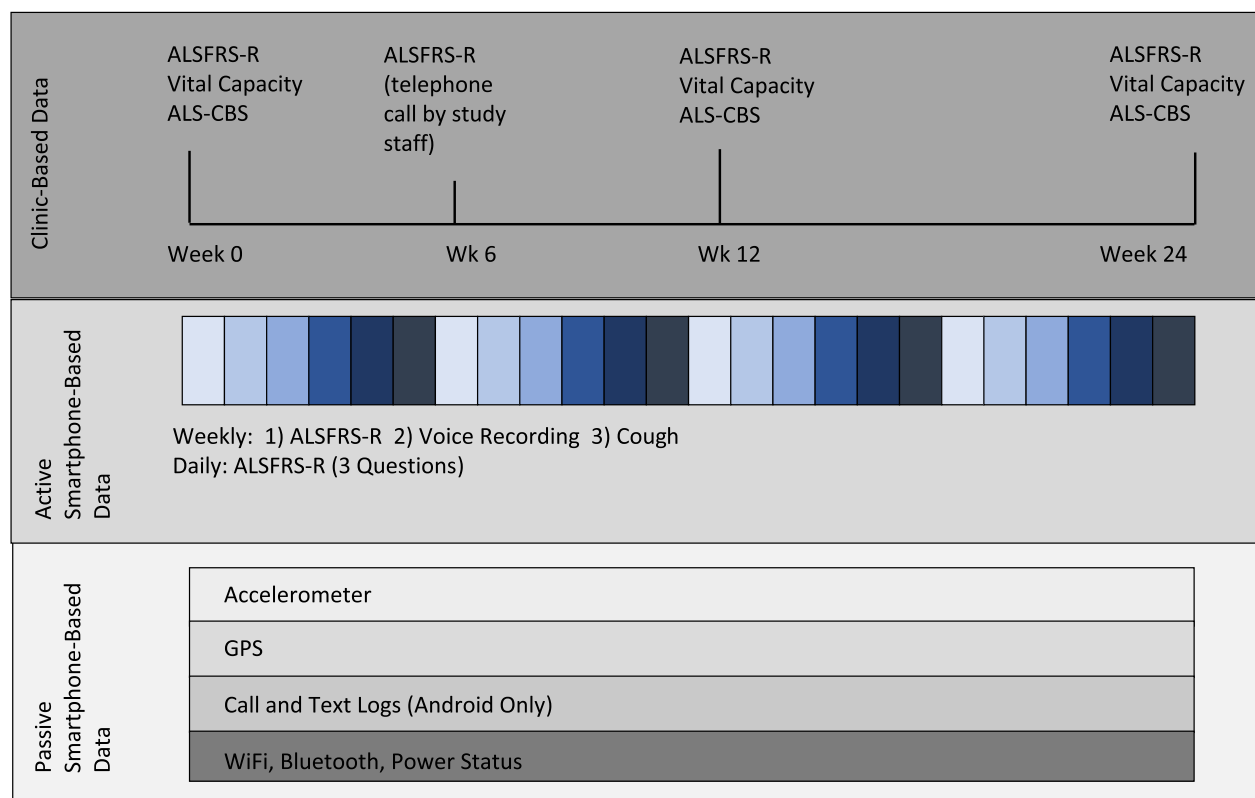
Figure 1. Study design: Data were collected in three data streams: (1) Clinic-based data collection at clinic visits and telephone calls; (2) Active smartphone-based data collection using questionnaires and audio recordings; (3) Passive smartphone-based collection of raw data from smartphone sensors and logs. ALSFRS-R: Revised ALS Functional Rating Scale; ALS-CBS: ALS Cognitive Behavioral Scale; GPS: Global Positioning System.

responses, voice recordings with scheduled prompts, and passive sensor and log data.

Approximately 6 weeks after enrollment, study personnel called the participant on the phone to obtain an ALSFRS-R. At approximately week 12, participants were seen in-person and study personnel collected clinic-based outcome measures. Participants were given the option to halt participation. For those who continued, participants were seen in clinic at approximately week 24 and study personnel collected in-person outcome measures. Participants could remove the Beiwe app from their phone at any time during the study (Fig. 1).

Clinic-based outcome measures

At clinic study visits, the ALS Functional Rating Scale (ALSFRS-R) was administered by study staff trained and certified by the Northeast ALS Consortium. The ALSFRS-R is a questionnaire made up of 12 questions in four domains (bulbar, fine motor, gross motor, respiratory) graded on a scale of 0–48 (each question 0–4; 0 is worst, 4 is normal function).[4] All 12 activities are relevant in

ALS, where speed of change in ALSFRS-R correlates with duration of survival.[5] The ALSFRS-R is validated for administration over the telephone[6] and it is used as the primary efficacy outcome in many ALS clinical trials.

The ALS Cognitive Behavioral Scale (ALS-CBS), used as an optional outcome measure, characterizes cognition and behavior in people with ALS and uses eight tasks for patients and a caregiver questionnaire to identify participants with frontotemporal dementia, which can co-occur in people with ALS.[7]

Vital Capacity (VC) was measured by trained and certified evaluators using an Easy One portable digital spirometer following standard techniques designed for use in ALS clinical trials.

Smartphone-based data collection

The Beiwe platform's web interface was used to configure active data collection tasks and to customize passive data collection. Study configuration data are stored in a single JSON configuration file that can be imported into a replication trial to validate findings. The contents of this file

specify all study settings for both active and passive data and are outlined in Table 1. The configuration file for this study is made available as a supplement (Data S1).

Active data collection tasks

Four days per week, participants were presented with notifications on the phone to answer three questions selected randomly without replacement from the ALSFRS-R. Using this method, over the course of each week, each of the 12 ALSFRS-R questions was presented once. In addition, the entire 12-question ALSFRS-R questionnaire was presented once weekly; the weekly surveys were analyzed. Daily surveys were removed after the first 10 participants completed the study due to informal participant feedback suggesting the frequency was burdensome.

Twice weekly, participants were asked to create an audio recording on the phone—once reading aloud a short paragraph (the Bamboo Passage) displayed on the phone screen,[8,9] and once recording a cough. Speech audio data were recorded by Beiwe using 44.1 kHz sampling rate without compression. Exploratory analysis of the cough spectrograms obtained is planned but has not yet been performed.

Speech audio recording processing

Speech recording files were preprocessed using Audacity 2.2.2. to reduce noise and to remove word insertions, word repetitions, filled pauses, and extraneous nonspeech sounds (e.g., throat clearing). The Speech Pause Analysis (SPA) software, a semiautomated MATLAB speech pause segmentation procedure,[8] was then used to extract the target speech and pause variables. SPA analysis calculates numerous speech and pause variables.[8,9] For the current study, we measured mean pause time, operationalized as the average duration of all pauses within a single recording of the Bamboo Passage, in seconds.[9] The mean pause duration was calculated by dividing the total pause duration in seconds by the number of pauses in the recording.

Table 1. Most relevant Beiwe settings for passive data collection.

Sensor type	On duration	Off duration
Accelerometer	10 sec	10 sec
GPS	60 sec	600 sec
Communication Metadata Log	Continuous	
Wi-Fi Usage Log	Single Log	300 sec
Power State	Continuous	

Additional settings can be found in Data S1.

Passive data collection

Passive data collected from the Beiwe app are divided into two groups: phone sensor data and phone logs. For phone sensor data, like GPS or accelerometer, sensor sampling alternates between two states, the on-cycle and the off-cycle. Data are collected continuously during the on-cycle and no data are collected during the off-cycle. Investigators are free to specify the duration of each cycle on Beiwe for each sensor separately. Phone logs, such as communication logs, are collected in full. No sampling was needed because their collection does not cause additional battery drain and their volume is smaller (Table 1). No actual content of communication is collected, and phone numbers in the communication logs are masked using a one-way hashing function.

Statistical methods

Demographics

Demographics, baseline ALS characteristics, clinic-based outcome measures, and baseline smartphone-based ALSFRS-R scores were summarized using descriptive statistics.

Outcome measures

A correlation between the clinic-based ALSFRS-R and smartphone-based ALSFRS-R was calculated by identifying measurement occasions that were as close to one another in time as possible. By temporally matching measures we hoped to make the most accurate comparison. In addition, a linear regression with difference between scoring approaches as the outcome and clinic-based ALSFRS-R as the only predictor was used to evaluate the agreement between clinic-based and smartphone-based scores across the spectrum of (clinic-based) ALSFRS-R scores. We then applied post hoc *t*-tests to compare individual item scores between clinic-based and smartphone-based ALSFRS-R. Speech measures were matched to clinic-based ALSFRS-R measures for a single time point analysis with the same temporal matching. A Pearson correlation was used for both comparisons.

To compare trajectories of smartphone-based total with clinic-based total, the smartphone data used were restricted to within half a month of the first and last clinic-based ALSFRS-R measurement for each participant to ensure that disease state was comparable across both approaches.

The ALSFRS-R trajectory was assessed separately for smartphone-based scores and clinic-based scores using a mixed model with ALSFRS-R total score as the outcome, a fixed effect for time (months from screening to

ALSFRS-R measurement), and a random intercept and slope for each participant with an unstructured covariance between effects. Models were compared to each other on magnitude and variation in the fixed effect for time estimate (change in ALSFRS-R total score over time) and on the variation in the random slope. The within- and between-subject variability from the smartphone model was also used to create a sample size calculation for a hypothetical future study.

A model was also created using a mixed effects linear regression with ALSFRS-R total score as the outcome, fixed effects for time and approach (smartphone based vs. clinic based) and an interaction between time and approach, and approach-specific random intercepts for each participant and approach-specific random slopes for each participant with an unstructured covariance between effects. The fixed effect for the interaction between approach and time in this model was used as an assessment of bias in the smartphone-based score compared to the clinic-based score. The correlation between the random slope term for smartphone-based data and the random slope term for clinic-based data was used to assess the correlation of approaches within each participant.

The trajectory of mean pause duration was estimated using a mixed effects model with mean pause as the outcome and time from first speech recording as the only predictor with a random slope and intercept for each participant with unstructured covariance.

All statistical analyses reported above were carried out in R (version 3.5). All tests were two-tailed with an alpha of 0.05.

Results

Demographic information

We consented to 23 participants, one passed away before data collection began. Participation in this pilot study occurred from 2016 to 2018. Baseline demographics and disease characteristics, including baseline ALSFRS-R range, are shown in Table 2.

ALSFRS-R correlation

The single time point correlation between clinic-based ALSFRS-R and smartphone-based ALSFRS-R was 0.93 (ALSFRS-R range: 19–47; $P < 0.001$) (Fig. 2). Mean self-report ALSFRS-R scores are 5% higher (2.4 points). Clinic-based ALSFRS-R and smartphone-based ALSFRS-R are equivalent at an ALSFRS-R total score at or above 39. Below a score of 39, for every 1-point decrease in the clinic-based score, there is a 1.75-point drop in the smartphone-based score ($P < 0.001$). Individual item

Table 2. Demographics.

Demographics	
Sex ($n = 23$)	
Male	16
Race ($n = 21$)	
Caucasian	19
Phone type ($n = 21$)	
Android OS	3
iOS	18
Onset location ($n = 21$)	
Bulbar	4
Upper Limb	8
Lower Limb	8
Trunk	1
Disease characteristics	
Mean disease duration at enrollment ($n = 21$)	31 mo (range 3–72)
Mean age ($n = 23$)	54
Mean diagnostic delay months ($n = 21$)	16
Mean baseline ALSFRS-R ($n = 22$)	34 (range 19–47)
Mean baseline vital capacity ($n = 20$)	67% predicted (range 26–109)

comparison of the 12 questions (each measured on a 0–4 scale) shows statistically significant differences between clinic-based and smartphone-based scores only for questions 3 (Swallowing; smartphone-based score 0.4 pts higher; $P = 0.031$) and 6 (Dressing and Hygiene; smartphone-based score 0.8 pts higher; $P < 0.001$).

ALSFRS-R decline

The average decline in the clinic-based ALSFRS-R was 0.43 (SD 0.14) points per month. The average decline in the smartphone-based ALSFRS-R was 0.53 (SD 0.10) points per month (Figs. 3 and 4). The combined model, however, did not have a significant effect for the interaction of time and approach ($P = 0.4$), suggesting that there was no significant bias between smartphone-based and clinic-based total score. The combined model estimates a random slope standard deviation of 0.26 for the smartphone-based approach and 0.56 for the clinic-based approach, suggesting that the smartphone-based score has lower within-person variation than the clinic-based score. The random slope terms were highly correlated ($r = 0.88$) suggesting no bias in within-person variation. Some smartphone-based ALSFRS-R data were missing, yet due to the frequent data collection, all participants had many more data points from smartphone-based ALSFRS-R than clinic-based ALSFRS-R (Fig. 5).

Given the observed within- and between-subject variability, a hypothetical clinical trial using monthly smartphone-based ALSFRS-R questionnaires would only

require 230 participants. This estimate assumes a trial duration of 6 months, treatment effect of 30% slowing of ALSFRS-R decline, and 80% power to detect the treatment effect at an alpha of 0.9. If ALSFRS-R was collected weekly, it would require only 200 participants, a 15% decrease. Incorporation of passive data would be expected to reduce the required sample size further.
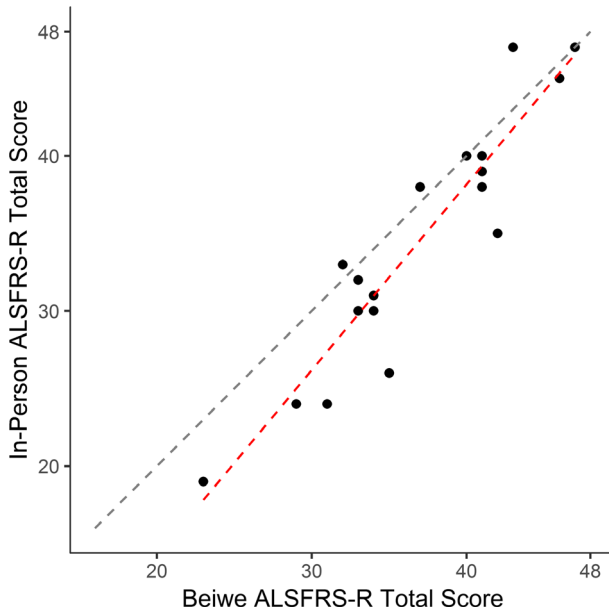

Figure 2. Correlation of clinic-based and smartphone-based ALSFRS-R is very high (Pearson *R* = 0.93). Scores diverge slightly at the lower end of the scale. The black dashed line represents a perfect correlation. The red dashed line represents the observed correlation.

Speech recording analysis

Participants successfully provided weekly speech recordings of the Bamboo Passage. There was an increase in mean pause duration (more impaired speech) over time in 10 out of 11 participants (range −0.02–0.05) (Fig. 6). The slope of the mean increase in pause time was 0.02 sec per month (0.009, NS). There was low correlation with the bulbar subdomain of the ALSFRS-R (−0.24 (NS)). The one participant with a decrease in mean pause time (improvement in speech) during the study reported no deficit in any bulbar or respiratory measure throughout the study except a 1-point decline (from 4 to 3) in swallowing at week 24.

Discussion

This smartphone-based digital phenotyping pilot study of people with ALS demonstrated that combining active and passive data collection is feasible and informative, and likely meaningful for ALS trial design.

To our knowledge, this is the first study to compare clinic-based and smartphone-based ALSFRS-R scores. The ALSFRS-R was developed to be a clinic-based tool. Our results suggest that self-report ALSFRS-R may be an equivalent or even better use of the tool.

This study demonstrates potential benefits of smartphone-based ALSFRS-R administration over the traditional clinic-based administration. First, clinic-based ALSFRS-R scores correlate highly with smartphone-based, self-reported ALSFRS-R scores at a single time point. Second, smartphone-based ALSFRS-R can be collected more
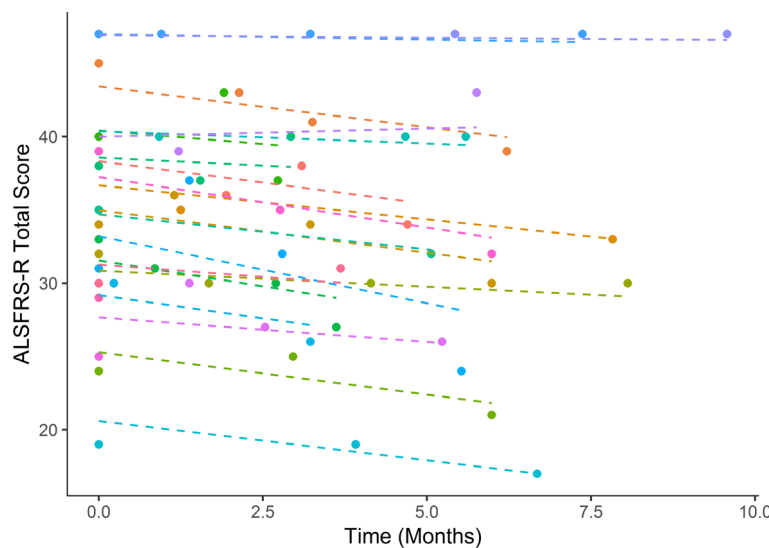

Figure 3. Clinic-based ALSFRS-R decline. Each dashed line represents one participant. Mean rate of decline was 0.43 (SD 0.14) points per month.
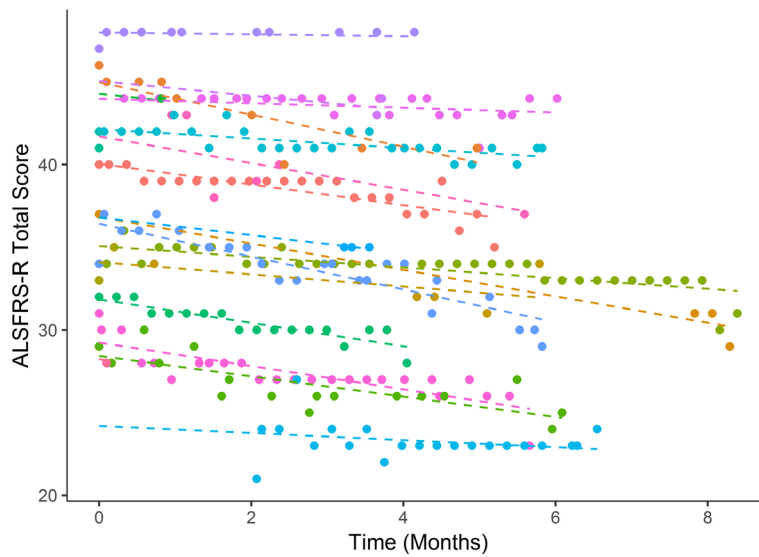
Figure 4. Smartphone-based ALSFRS-R decline. Mean rate of decline was 0.53 (SD 0.10) points per month.
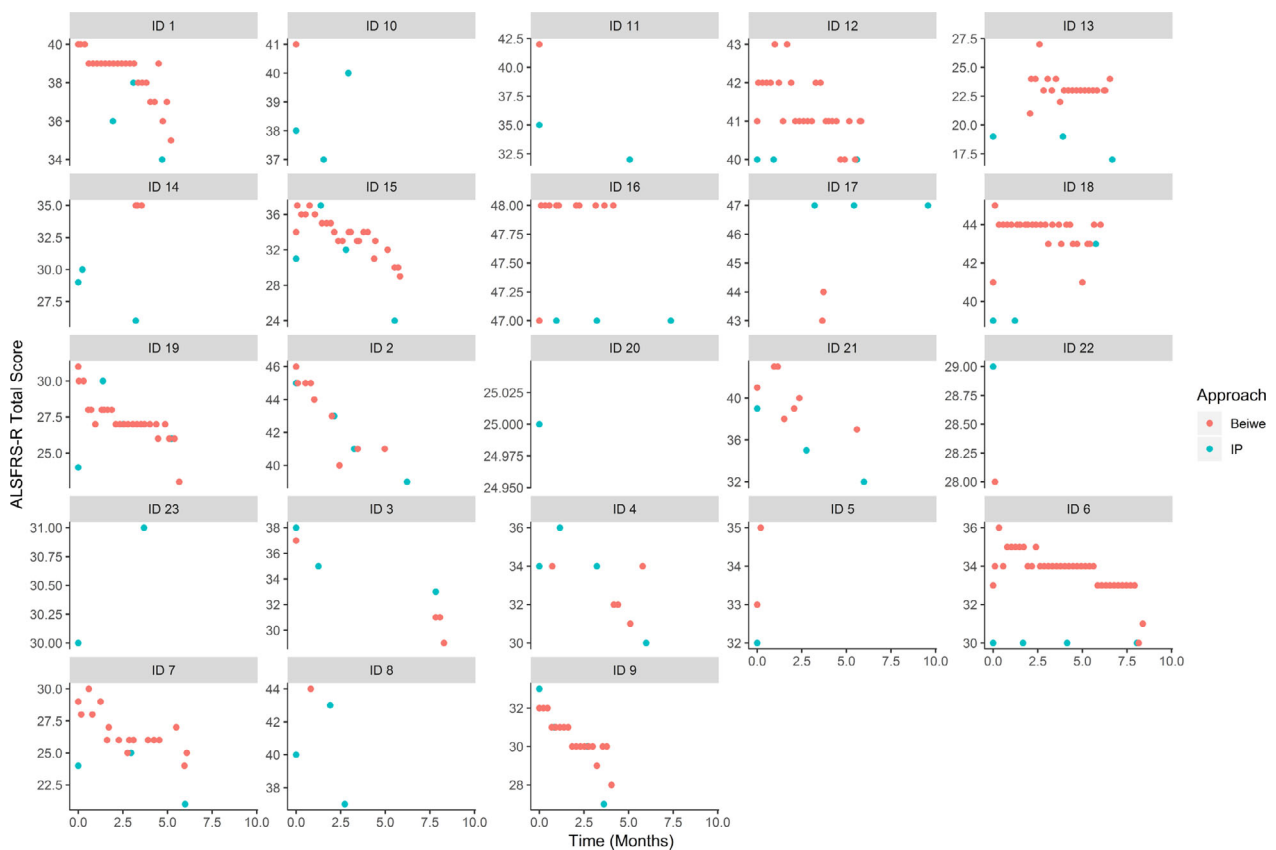

Figure 5. Correlation of clinic-based (teal circles) and smartphone-based (red circles) ALSFRS-R. Each panel corresponds to one participant.

frequently, increasing statistical power for a treatment trial. This simple change in frequency of administration could lead to a 15% reduction in required sample size in

a 6-month trial, a substantial improvement. Third, smartphone-based and clinic-based ALSFRS-R slopes are comparable and there is a lower within-person variability in
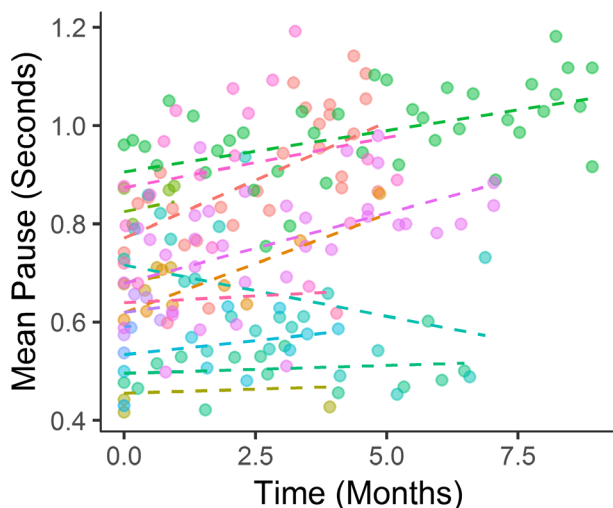
Figure 6. Mean pause (total pause duration in a speech recording divided by the number of pauses) increased over time in all participants except one.

smartphone-based ALSFRS-R, potentially increasing statistical power even more.

There was no statistically significant difference between clinic-based and smartphone-based ALSFRS-R scores, but we did find a lower correlation at the lower end of the scale. In addition, discrepancy varied by question, with the greatest discrepancies seen in questions 3 (Swallowing) and 6 (Dressing and Hygiene). It is possible that the language of these items, particularly at the lower end of function, is unclear to patients. The use of previously published patient-friendly wording of the ALSFRS-R might reduce these discrepancies.[10]

At the same time, not all differences between clinic-based measurements and smartphone-based measurements represent error in self-report. In a previous study using the Beiwe platform in a psychiatric population, more participants acknowledged feelings of suicidality on smartphone-based questionnaires than their clinic-based equivalent,[1,2] suggesting smartphone-based surveys might elicit more truthful responses than clinic-based surveys.

Quantified analysis of motor speech decline provides insight into disease progression in the bulbar region. In this preliminary analysis, we show that speech recordings can be collected frequently at home in a natural setting. Further analysis of speech recording features will identify the optimal features for detecting early changes in bulbar function and for following changes over time within individuals. To our knowledge, only one ALS trial has used quantified motor speech decline as a trial outcome, and this trial demonstrated a statistically significant effect of the study drug.[11] Smartphone-based speech recording could further increase statistical power for detecting

bulbar motor involvement by making it easier for individuals to participate in trials and by increasing the number of representative speech samples for each participant, just as we have demonstrated for smartphone-based ALSFRS-R administration.

We plan to conduct more in-depth analyses of both the speech recording and passive smartphone data collected in this pilot study. We have also incorporated Beiwe into an ongoing clinical trial of inosine for people with ALS and will initiate a larger, multicenter observational study using Beiwe. Already it seems clear that smartphone-based data collection in people with ALS is feasible, informative, and can add statistical power in clinical trials.

Acknowledgments

Conflict of Interest

None declared.

References

1. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016;3:e16.

2. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016;41:1691–1696.

3. Onnela J-P. Available from: https://github.com/onnela-lab.

4. Cedarbaum JM, Stambler N, Malta E, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). J Neurol Sci 1999;169:13–21.

5. Kimura F, Fujimura C, Ishida S, et al. Progression rate of ALSFRS-R at time of diagnosis predicts survival time in ALS. Neurology 2006;66:265–267.

6. Kasarskis EJ, Dempsey-Hall L, Thompson MM, et al. Rating the severity of ALS by caregivers over the telephone using the ALSFRS-R. Amyotroph Lateral Scler Other Motor Neuron Disord 2005;6:50–54.

7. Woolley SC, York MK, Moore DH, et al. Detecting frontotemporal dysfunction in ALS: utility of the ALS Cognitive Behavioral Screen (ALS-CBS). Amyotroph Lateral Scler 2010;11:303–311.

8. Green JR, Beukelman DR, Ball LJ. Algorithmic estimation of pauses in extended speech samples of dysarthric and typical speech. J Med Speech Lang Pathol 2004;12:149–154.

9. Yunusova Y, Graham NL, Shellikeri S, et al. Profiling speech and pausing in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). PLoS ONE 2016;11: e0147573.

10. Montes J, Levy G, Albert S, et al. Development and evaluation of a self-administered version of the ALSFRS-R. Neurology 2006;67:1294–1296.

11. Green JR, Allison KM, Cordella C, et al. Additional evidence for a therapeutic effect of dextromethorphan/quinidine on bulbar motor function in patients with amyotrophic lateral sclerosis: a quantitative speech analysis. Br J Clin Pharmacol 2018;84:2849–2856.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1. Settings for the Beiwe platform are recorded in this file and could be used to replicate the exact data collection parameters from this study in a future study.