

# Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration

ISSN: 2167-8421 (Print) 2167-9223 (Online) Journal homepage: [www.tandfonline.com/journals/iafd20](http://www.tandfonline.com/journals/iafd20)

## Estimation of forced vital capacity using speech acoustics in patients with ALS

Gabriela M. Stegmann, Shira Hahn, Cayla J. Duncan, Seward B. Rutkove, Julie Liss, Jeremy M. Shefner & Visar Berisha

**To cite this article:** Gabriela M. Stegmann, Shira Hahn, Cayla J. Duncan, Seward B. Rutkove, Julie Liss, Jeremy M. Shefner & Visar Berisha (2021) Estimation of forced vital capacity using speech acoustics in patients with ALS, *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 22:sup1, 14-21, DOI: [10.1080/21678421.2020.1866013](https://doi.org/10.1080/21678421.2020.1866013)

**To link to this article:** <https://doi.org/10.1080/21678421.2020.1866013>



© 2021 The Author(s). Published by Informa UK Limited trading as Taylor & Francis Group on behalf of the Montreal Neurological Institute-Hospital



Published online: 04 Aug 2021.



[Submit your article to this journal](#)



Article views: 3660



[View related articles](#)



[View Crossmark data](#)



Citing articles: 20 [View citing articles](#)

## RESEARCH ARTICLE

# Estimation of forced vital capacity using speech acoustics in patients with ALS

GABRIELA M. STEGMANN<sup>1,2</sup>, SHIRA HAHN<sup>1,2</sup>, CAYLA J. DUNCAN<sup>3</sup>, SEWARD B. RUTKOV<sup>4</sup>, JULIE LISS<sup>1,2</sup>, JEREMY M. SHEFNER<sup>3,5,6</sup> & VISAR BERISHA<sup>1,2</sup>

<sup>1</sup>Department of Speech Pathology, Arizona State University, Phoenix, AZ, USA, <sup>2</sup>Aural Analytics, Scottsdale, AZ, USA, <sup>3</sup>Barrow Neurological Institute, Phoenix, AZ, USA, <sup>4</sup>Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA, <sup>5</sup>Creighton University College of Medicine, Phoenix, AZ, USA, <sup>6</sup>Department of Neurology, University of Arizona College of Medicine, Phoenix, AZ, USA

## Abstract

In this study, we present and provide validation data for a tool that predicts forced vital capacity (FVC) from speech acoustics collected remotely via a mobile app without the need for any additional equipment (e.g. a spirometer). We trained a machine learning model on a sample of healthy participants and participants with amyotrophic lateral sclerosis (ALS) to learn a mapping from speech acoustics to FVC and used this model to predict FVC values in a new sample from a different study of participants with ALS. We further evaluated the cross-sectional accuracy of the model and its sensitivity to within-subject change in FVC. We found that the predicted and observed FVC values in the test sample had a correlation coefficient of .80 and mean absolute error between .54 L and .58 L (18.5% to 19.5%). In addition, we found that the model was able to detect longitudinal decline in FVC in the test sample, although to a lesser extent than the observed FVC values measured using a spirometer, and was highly repeatable (ICC = 0.92–0.94), although to a lesser extent than the actual FVC (ICC = .97). These results suggest that sustained phonation may be a useful surrogate for VC in both research and clinical environments.

**Keywords:** *Clinical trial, respiratory function, speech analysis, biomarker, ventilation*

## Introduction

Optimal management of patients with amyotrophic lateral sclerosis (ALS) and other neuromuscular diseases requires the ability to accurately assess respiratory function. Timely assessment of pulmonary function is important in gauging prognosis and instituting care such as noninvasive ventilatory support. Approximately half of ALS patients in the United States receive care in multidisciplinary clinics where vital capacity and other pulmonary function studies are routinely performed. Telemedicine is quickly emerging as an appealing alternative to in-clinic visits as it reduces patient burden and allows for more frequent evaluations. While existing telemedicine solutions allow for efficient real-time communication between patients and doctors, remote objective assessment of patients' pulmonary

function requires the use of specialized hardware, namely spirometers. This prevents widespread use and has resulted in a recognized need that innovative solutions are required for low-cost, low-burden, remote assessment of pulmonary function, particularly for vulnerable populations (1).

There are several studies that have collected vital capacity information remotely (2); however, they all require measurement using an at-home spirometer. As an alternative, there is growing evidence in the literature that speech-based tasks such as single breath count and maximum sustained phonation provide moderate correlations with vital capacity in healthy participants (3,4) and those with asthma or COPD (5). This is an appealing alternative to spirometry-based assessment as speech can be elicited and measured using

Correspondence: Jeremy M. Shefner, Barrow Neurological Institute, 240 West Thomas Road, Phoenix, AZ 85013, USA. E-mail: [Jeremy.shefner@dignityhealth.org](mailto:Jeremy.shefner@dignityhealth.org)

(Received 24 October 2020; revised 23 November 2020; Accepted 8 December 2020)

ISSN print/ISSN online © 2021 The Author(s). Published by Informa UK Limited trading as Taylor & Francis Group on behalf of the Montreal Neurological Institute-Hospital

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI: [10.1080/21678421.2020.1866013](https://doi.org/10.1080/21678421.2020.1866013)

a mobile app without the need for specialized equipment. In this paper we present a new tool for predicting FVC and systematically investigate its performance on participants with ALS. Here we report findings derived from two separate observational, longitudinal studies where participants with ALS had longitudinal measurements of both FVC (measured via spirometer, either by the patient at home or by a trained evaluator in clinic) and sustained phonations obtained at home using an app installed on the patient's mobile device.

## Methods

### *Data collection*

Data was collected from two separate studies. The first study was used to train a model which could predict FVC, and a separate study was used to test the accuracy of the prediction model. The two samples are described below. For both studies, approval was granted by the Institutional Review Board of St. Joseph's Hospital and Medical Center, and all subjects signed an informed consent document.

**Training sample.** The training sample was from ALS at Home (6,7), an observational, longitudinal study that was conducted entirely remotely. This study included healthy participants and participants with amyotrophic lateral sclerosis (ALS). Participants were recruited and assessed from home. For the first three months, participants were requested to provide daily speech samples and spirometry measures, as well as measures of grip strength, general activity, electrical impedance myography, and a weekly self-assessment of function using the ALS Functional Rating Scale - Revised (ALSFRRS-R). Subsequently, participants were requested to perform the same battery twice per week for an additional six months. Participants were allowed to receive assistance from their caregivers if needed. The FVC and sustained phonation tasks were performed from home. The original training sample consisted of 110 participants. 55 participants did not provide FVC data (the outcome of interest) and 46 participants did not provide height data (a necessary predictor). After removing all participants with no overlapping

FVC and phonation data or no height provided, the final training sample consisted of 39 participants, each one measured repeatedly throughout the study such that there were 1,971 observations with complete data.

**Test sample.** The test sample was obtained from an ongoing observational longitudinal study of ALS patients that is currently in progress, whose purpose is to assess speech and language changes in ALS patients with and without symptoms of cognitive dysfunction. Every three months, participants attended a clinic where FVC was measured and a speech sample obtained (including maximum-effort sustained phonation) with clinician supervision. Participants also provided weekly speech samples that included a maximum-effort sustained phonation from home. There was a total of 25 participants with 47 FVC measurements and 578 sustained phonation measurements. Table 1 shows the descriptive statistics for the final training and test samples. To minimize participant burden in both studies participants were only asked to provide one attempt at MPT, as multiple other measures were being assessed at each session.

### *Speech collection and analysis*

For both studies, participants downloaded a mobile app onto a device (smartphone or tablet) which was used for recording an ambient noise sample and a maximum-effort sustained phonation task. Prior to vocal tasks being performed, a sample of ambient noise was recorded, and this was used to filter out signals during the task that were not participant derived. Participants were instructed to take a deep breath and say "ahh" for as long as possible until they ran out of breath. The recorded speech sample was saved as a .wav file and sent to a secure, HIPAA-compliant cloud-based repository for storage. A speech analysis algorithm used the ambient noise sample and the sustained phonation to automatically assess the maximum phonation time (MPT)—the length of the sustained phonation. This was done by calculating the time from phonation onset to phonation offset, as determined by a speech activity detector. The automatic MPT algorithm is accurate to within a mean absolute error of 0.01 seconds, as

Table 1. Descriptive demographic statistics for training and test samples.

Description	Training Sample	Test Sample
Number of Participants	33 ALS; 6 HC	25 ALS
Total Observations	1,971 (overlapping FVC and sus. phon.)	47 FVC obs. 578 sus. phon.
Gender	15 F; 24 M	8 F; 17 M
Age (mean, SD)	58.1 (SD = 9.4)	65.4 (SD = 11.8)
FVC (L) (mean, SD)	3.51 (SD = 1.15)	3.38 (SD = 1.31)
ALSFRRS-R total (in ALS participants)	37.1 (SD = 7.3)	36.9 (SD = 6.0)
Height (cm)	172.0 (8.6)	173.8 (SD = 11.5)

determined by a comparison of automatically calculated MPT values to ground truth MPT values computed from 100 randomly selected speech files, where speech onset and offset were manually labeled by trained annotators. Trials were excluded if no perceptible patient phonation could be detected.

#### *FVC collection*

In the training sample, FVC measurements were performed from home. Participants were provided an AirSmart Spirometer (Nuvoair AB, Stockholm, Sweden). For each session, participants were asked to perform three FVC maneuvers, with the highest of the three uploaded to Air Smart’s health cloud, and transferred to REDCap database (6).

In the test set, patients submitted speech and language samples via the app weekly from home and at their routine clinic visits. Patients were seen every three months as part of their clinical care; as part of standard of care, FVC was obtained by a licensed respiratory therapist or an evaluator trained and certified by the NEALS consortium, with 3–5 attempts recorded at each visit.

#### *FVC prediction model*

A machine learning (ML) model was trained to predict FVC based on the at-home data. Several acoustic features and demographic characteristics were considered, including MPT, measures of pitch, loudness, and vocal quality extracted from the sustained phonation, and age, height, gender, and weight. A mixed-effects framework was used to account for the repeated measurements per participant. To separate the between-person effects and within-person effects, each feature extracted from the phonation was disaggregated such that each participant would have a mean for each predictor and a deviation from the mean for each observation, following the within-person effects disaggregation method described in (8).

Both linear and nonlinear models consisting of different sets of variables were tested. The performance of each model was evaluated using leave-one-participant-out cross-validation on the training data, in a manner similar to that used in (9). The model was estimated on a training sample consisting of all participants minus one, and the outcome was predicted on the participant that was left out of the training sample using the estimated model. This process was repeated leaving out one participant at a time. The performance of each model was evaluated using the mean absolute error (MAE, described below) between the predicted and observed FVC values using the out-of-sample predictions (the estimates obtained in each participant that was left out while training set for the

model) (9). The final model was a linear model which included age, height, and MPT as features.

#### *Evaluation of cross-sectional prediction accuracy*

Once the final training model was obtained, we tested its accuracy on the test sample described above.

Using the features and parameters estimated from the final training model, a predicted FVC measure was obtained for each observation in the test set. Prediction accuracy was evaluated using the mean absolute error (MAE) between the observed FVC measures and the FVC measures predicted according to the model:

$$MAE = \sum_{i=1}^n \sum_{j=1}^{t_i} |FVC_{ij} - \widehat{FVC}_{ij}|$$

where  $i$  is the  $i$ -th participant,  $j$  is the  $j$ -th observation for the  $i$ -th participant,  $FVC$  is the observed FVC value, and  $\widehat{FVC}$  is the predicted FVC value. The MAE is interpreted in the same units as the original outcome, which in this case is FVC L. Lower MAE scores indicated better prediction accuracy.

#### *Evaluation of longitudinal change*

After evaluating the prediction accuracy of the model, we evaluated how the observed FVC and the predicted FVC tracked with longitudinal change. To evaluate the longitudinal change, we used a growth curve model (GCM) (10). A GCM is a mixed-effects model where the dependent variable is the outcome of interest and the primary predictor is the time variable. The following GCM was used:

$$FVC_{ij} = b_{0i} + b_{1i} \cdot t_{ij} + e_{ij}$$

where  $FVC_{ij}$  is the FVC value for individual  $i$  at time  $j$ ;  $b_{0i}$  is the intercept (e.g. the expected FVC measure when  $t_{ij} = 0$ ) for participant  $i$ , which follows a normal distribution with mean intercept  $\beta_0$  (fixed effect) and a standard deviation;  $b_{1i}$  is the mean slope (i.e. rate of change) for all individuals;  $t_{ij}$  is the value of the time variable (e.g. number of days since enrollment) for individual  $i$  at time  $j$ ; and  $e_{ij}$  is the residual term for individual  $i$  at time  $j$ . Although we attempted to allow for a unique slope for all individuals by allowing  $b_{1i}$  to have a random term (i.e. standard deviation), participants had very few observed FVC measures and the model did not converge. Therefore, the final model (which did converge) allowed for random intercepts but no random slopes.

We fit two separate GCMs: one for the observed FVC measures and one for the FVC measures predicted by the model. The time variable was the number of days since enrollment in the study. Prediction accuracy of the final model

on the training sample was evaluated using leave-one-participant-out cross-validation. All correlations reported were adjusted for the repeated measurements per participant (11). We estimated

the repeatability of the MPT prediction using the intra-class correlation (ICC), the within-person standard error of measurement (SEM), and the within-person coefficient of variation (CV) (12).

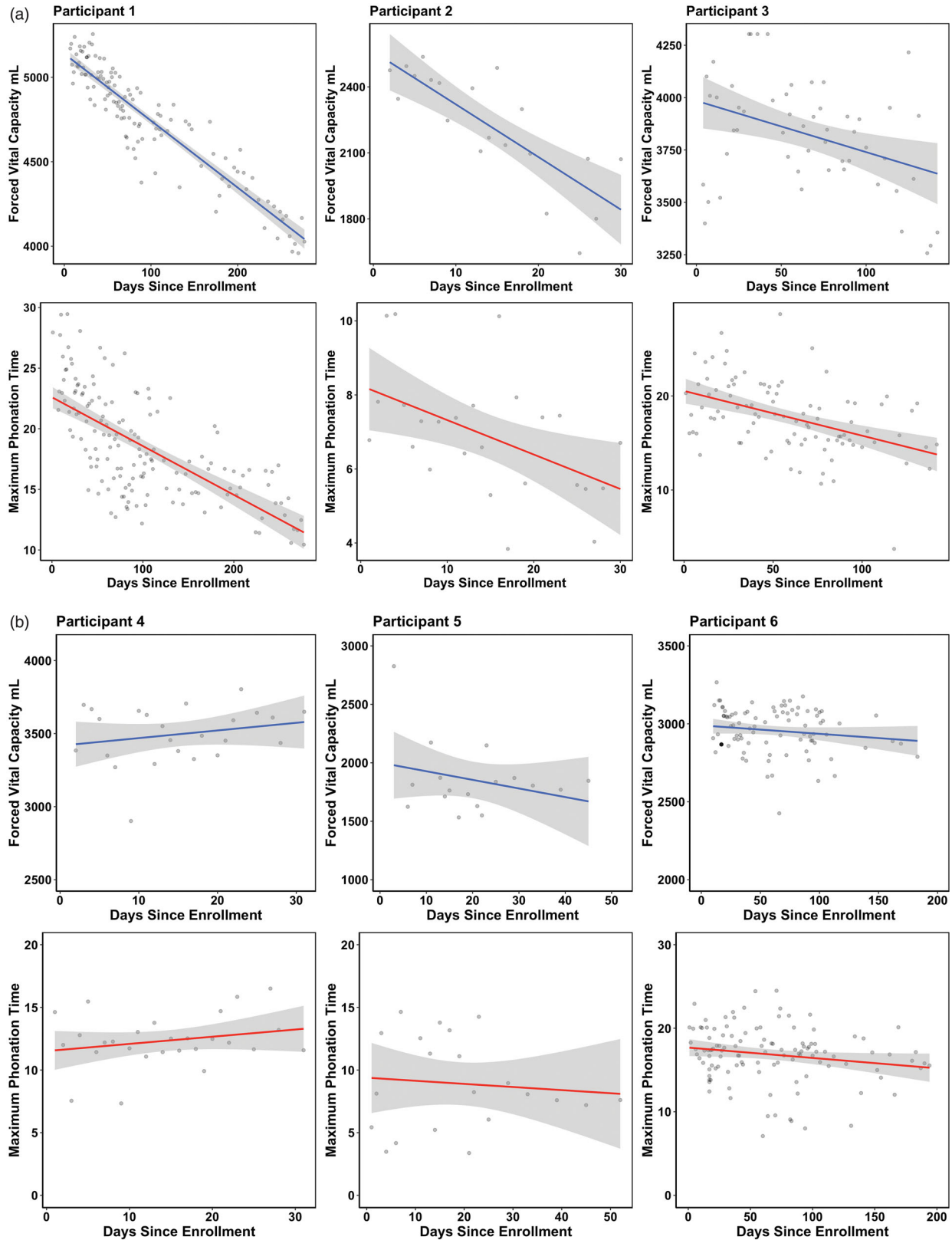


Figure 1. Examples of declining (a) and stable (b) FVC and MPT trajectories. The gray points show the observed participants' FVC and MPT values. The blue solid line shows the mean trajectory (obtained from the regression equation) for the FVC for each participant. The solid red line shows the mean trajectory for the MPT for each participant. The gray shading around the lines shows the confidence band for the mean trajectory.



## Results

### *Evaluation of cross-sectional prediction accuracy*

Figure 1 displays individual trajectories of participants from the training set for FVC (blue) and MPT. We found that the MAE was .47 L (relative MAE = 14%) and the correlation coefficient  $r$  was .72. We then used the trained model to obtain FVC predictions on the test sample and evaluated the performance on that sample. In the test sample, the MAE was .58 L (relative MAE = 19.5%) and the correlation coefficient  $r$  was .80, meaning that on average, the predicted FVC deviated .58 L from the observed FVC in the test sample. Finally, we evaluated the prediction accuracy using the average of 3 and 5 MPT measurements for prediction (i.e. the average of the observation overlapping with FVC, the one or two before, and the one or two after), and found that prediction

Table 2. Model fit for both training sets, expressed as Mean Absolute Error (MAE) and correlation coefficient ( $r$ ).

Model	MAE (L)	Relative MAE	$r$
Training set: leave-one-participant-out cross-validation	.47	14%	.72
Test set: predict FVC using single observation	.58	19.5%	.80
Test set: predict FVC using the three nearest observations	.56	18.9%	.80
Test set: predict FVC using the five nearest observations	.54	18.5%	.80

accuracy increased when more MPT measurements were used. The model fit results are shown in Table 2, including the MAE, relative MAE, and  $r$ . Figure 2 shows predicted and observed FVC in a scatterplot using the best model, consisting of the 5 closest MPT measurements.

### *Evaluation of longitudinal change*

We fit a GCM to the observed and predicted FVC values in the training sample (via cross-validation) and test sample, and we evaluated the longitudinal slopes (mean rate of change) for both sample sets. Table 3 shows the GCM parameters for the observed and predicted FVC models using both the cross-validated training data and the test data. The fixed-effects intercepts indicate the expected intercept (expected FVC at the start of the study), the fixed-effects slopes indicate the expected rates of change (expected decline FVC in L per month), the intercepts standard deviation indicates how much participants varied in their unique intercepts (how different participants were at the start of the study), and the residual standard deviation is how much each observation deviated from each participant's unique trajectory. The slopes had fixed effects but not random effects. The final models reported in this paper all converged appropriately. Both GCMs yielded significantly negative slopes, indicating that both observed and predicted FVC were declining throughout the study.

Figure 3 shows the observed data for the FVC and predicted FVC side-by-side. For clearer visualization of the trajectories, only those participants

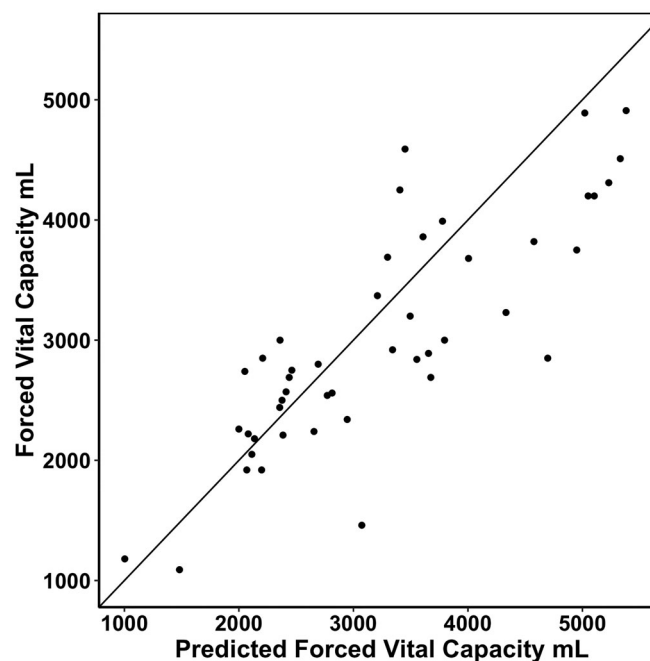


Figure 2. Observed and predicted FVC on the test data. Each point shows each FVC observed and predicted value. The diagonal line shows where the region where the observed and predicted values are identical. The closer the points are to the diagonal line, the lower the prediction error is.

Table 3. Parameters for the Growth Curve Models (GCMs) for both training sample and test sample.

Parameter	Test Data Model: Observed FVC	Test Data Model: Predicted FVC	Training Data Model: Observed FVC	Training Data Model: Predicted FVC
	Mean (standard error)	Mean (standard error)	Mean (standard error)	Mean (standard error)
Fixed effects				
Intercept (in Liters)	3.28 (.24)	3.27 (.23)	3.64 (.18)	3.61 (.15)
Slope (in L/month)	-.037 (.01)	-.017 (.004)	-.027 (.005)	-.024 (.006)
Random effects				
Intercepts standard deviation	1.15	1.13	1.14	.94
Residual standard deviation	.17	.33	.19	.24

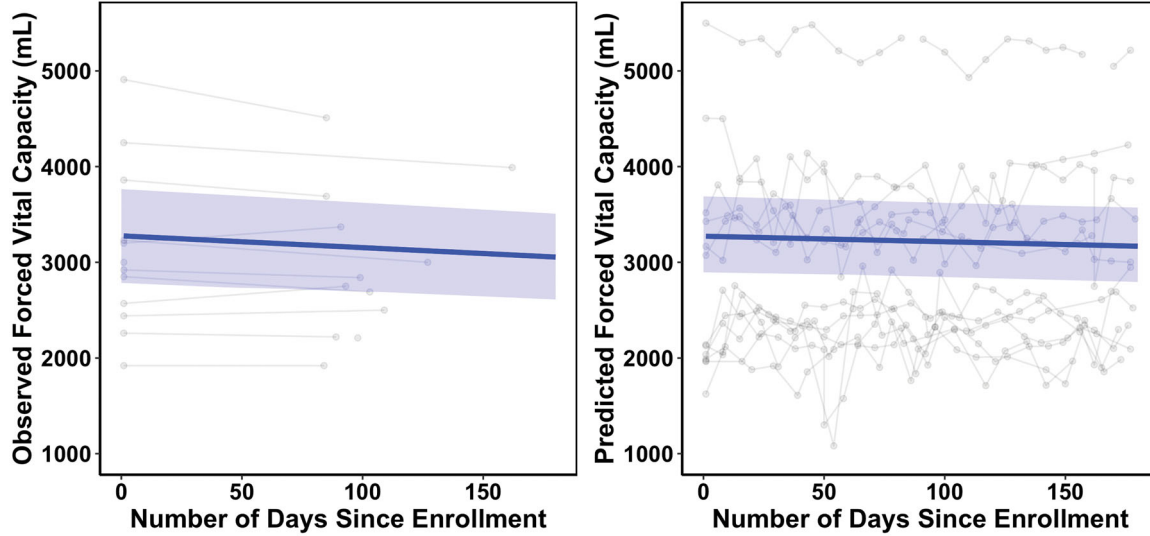


Figure 3. Observed and predicted FVC longitudinal trajectories. The gray points show the observed FVC data (left) and the predicted FVC data according to the model (right). Each participant's data are connected by the gray lines. The solid blue line and the blue shade show the expected longitudinal trajectory and confidence band according to the growth curve model.

with at least 15 phonation measurements were included in the plots. However, the full sample was used for the analyses. The dark blue lines show the predicted trajectory according to the GCM and the blue shades show the 95% confidence band, which was estimated using the predicted population interval method (13). As shown in the plots, both predicted and observed FVC values had very similar intercepts and declining trajectories; however, the predicted FVC values declined at a slower rate than the observed FVC values.

#### Repeatability of the predicted FVC measurement

The test-retest repeatability scores were computed for observed FVC and predicted FVC values on the training and test data. The ICC ranges from 0 to 1, where higher scores indicate higher repeatability. SEM is the within-person standard deviation, and it is expressed in the observed units of FVC (L), such that lower values indicate lower variability (and therefore higher repeatability). The CV is the within-person variability (standard deviation) divided by the mean of the data, and it is expressed as a percentage, such that lower values

Table 4. Reproducibility of model using Interclass Correlations (ICCs), Standard Error of the Mean, (SEM) and Coefficient of Variation (CV).

Model	ICC	SEM	CV
Observed FVC in Training Data	0.97	0.19	6%
Predicted FVC in Training Data	0.94	0.24	9%
Observed FVC in Test Data	0.97	0.22	6%
Predicted FVC in Test Data	0.92	0.34	15%

indicate lower variability (and therefore higher repeatability). Table 4 shows the repeatability scores for the observed and predicted FVC in the training and test sets. We found that the repeatability in the predicted FVC was overall lower than the repeatability in the observed FVC, but still appropriate for clinical applications (14,15).

#### Discussion

A mixed-effects machine learning model was constructed to predict FVC from height, age, and MPT. Cross-sectionally, this model had a maximum out-of-sample accuracy of .54 L MAE (18.5% relative MAE) with a correlation between

the predicted and observed FVC values of  $r = .80$ . To predict longitudinal change, a growth curve model was fit to observed and predicted FVC. The slope of the predicted FVC was slightly less steep than the slope of the observed FVC for the test sample. There are two possible explanations for this. First, the model was trained using at-home spirometry measures whereas the test sample used in-clinic spirometry measures. Second, participants performed at-home spirometry without the guidance of a clinician, whereas the in-clinic spirometry was administered by a respiratory therapist according to standard protocol.

The repeatability of the FVC prediction was quite good, though slightly lower than the repeatability of the observed FVCs. This was a result of the lower reliability of the MPT measurements, also observed in other studies (16). Several studies have analyzed how to elicit maximum-performance sustained phonation in other contexts and have suggested that modeling and repeat performance of the sustained phonation task increases MPT (17) and improves reliability (18). Future studies that aim to assess FVC via MPT would benefit from modeling the sustained phonation via training videos and then repeating performance of the task for each session and taking the maximum among the tasks.

Unlike a standard FVC test, the sustained phonation is modulated by both the vital capacity and the valving of the column of exhaled air by the vocal folds. Thus, maximum phonation time is impacted both by phonatory function and respiratory function. This is an important consideration in the present study, as vocal quality may change over time in ALS, especially in the case of bulbar-onset (19). To explore the relative contributions of phonatory and respiratory function to MPT, we fit two models using FVC as a proxy for respiration and cepstral peak prominence (CPP) as a proxy for phonation quality (20), and for each model, we estimated the amount of variation explained by the predictor using the  $R^2$  for mixed-effects models as described in (11). In the first model, we predicted MPT based on FVC and found that  $R^2 = .24$ . Then we fit a model where we predicted MPT based on CPP alone and found that  $R^2 = .01$ . Therefore, the variability in MPT was moderately influenced by respiration, but only mildly influenced by the vocal quality of the phonatory function. The strong association between FVC and predicted FVC in our results also support MPT as a measure for respiratory function, with only minimal impact from the quality of the phonation for these participants. It would be interesting to determine whether specific clinical characteristics such as bulbar disease burden contributed to reliability and/or predictive accuracy, but patient numbers were not sufficient to assess this.

MPT is only one of a number of tasks that might be predictive of VC; it will be up to future studies to determine whether MPT can function in a useful manner and if other tasks may provide equivalent or improved predictive capacity. The extent to which MPT might serve as a useful outcome measure in clinical trials is an important question. The natural history cohorts that form the basis of this study were not selected with clinical trial inclusion criteria in mind; future studies in cohorts more representative of the clinical trial population will help to determine how this measure functions in that environment. Further experience in the clinical setting will also help determine the extent to which MPT can serve as a clinically useful surrogate in clinical situations where VC cannot be obtained, either because a visit is being conducted remotely or if the procedure is deemed a risk for any reason.

## Conclusion

In this study, we showed that it was possible to assess respiratory function using a maximum-performance phonation task. This could be done remotely and by using a phone, without the need for specialized equipment. We found that predicted FVC values mapped onto observed FVC measurements on a new sample that was not used for training the model. The GCMs showed that the predicted FVC tracked longitudinally, although to a lesser extent than the actual FVC measurements. We also found that the test-retest reliability was lower than the actual FVC, but the reliability was still commensurate with other commonly used outcome measures in ALS (14,15).

## Declaration of interest

Dr. Visar Berisha is an Associate Professor at Arizona State University. He is a co-founder of Aural Analytics. Dr. Julie Liss is a Professor and Associate Dean at Arizona State University. She is a co-founder of Aural Analytics. Dr. Jeremy Shefner is the Kemper and Ethel Marley Professor and Chair of Neurology at the Barrow Neurological Institute. He is a scientific advisor to Aural Analytics.

## Funding

This work was supported by NIH SBIR [1R43DC017625-01], NSF SBIR [1853247], NIH R01 [5R01DC006859-13], ALS Finding a Cure Grant, and a grant from the Dignity Health ASU Research Collaboration.



## References

1. Hull J, Lloyd J, Cooper B. Lung function test in the COVID-19 endemic. *Lancet Respir Med*. 2020;8:666–7.
2. Geronimo A, Simmons Z. Evaluation of remote pulmonary function test in motor neuron disease. *Amyotroph Lateral Scler Frontotemporal Degener*. 2019; 20:348–6.
3. Lima Dcbb, Palmeira AC, Costa EC, Mesquita FODS, Andrade FMDD, Correia Júnior MADV. Correlation between slow vital capacity and the maximum phonation time in healthy adults. *Rev Cefac*. 2014;16:592–7.
4. Palaniyandi A, Natarajan M, Chockalingam A, Karthick R. Even a single breath counts. *J Dent Med Sci*. 2017;16:70–2.
5. Chun KS, Nathan V, Vatanparvar K, Nemati E, Rahman MM, Blackstock E, Kuang J. Towards passive assessment of pulmonary function from natural speech recorded using a mobile phone. *Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 2020:1–10.
6. Rutkove SB, Qi K, Shelton K, Liss J, Berisha V, Shefner JM. ALS longitudinal studies with frequent data collection at home: study design and baseline data. *Amyotroph Lateral Scler Frontotemporal Degener*. 2019;20:61–7.
7. Rutkove SB, Narayanaswami P, Berisha V, Liss J, Hahn S, Shelton K, et al. Improved ALS clinical trials through frequent at-home self-assessment: a proof of concept study. *Ann Clin Transl Neurol*. 2020;7:1148–57.
8. Curran PJ, Bauer DJ. The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu Rev Psychol*. 2011;62:583–619.
9. Yang Y, Huang S. Suitability of five cross validation methods for performance evaluation of nonlinear mixed-effects forest models—a case study. *Forestry* 2014; 87:654–62.
10. Grimm KJ, Ram N, Estabrook R. Growth modeling: structural equation and multilevel modeling approaches. New York: Guilford; 2017.
11. Lorah J. Effect size measures for multilevel models: definition, interpretation, and TIMSS example. *Large Scale Assess Educ*. 2018;6:1–11.
12. Bland M. An introduction to medical statistics. Oxford, UK: Oxford University Press; 2015.
13. Bolker BM. Optimization and all that. In: *Ecological models and data in R*. Princeton, NJ: Princeton University Press; 2019.
14. Fleiss JL. The design and analysis of clinical experiments. New York, NY: Wiley; 1999.
15. Portney L, Watkins M. Foundations of clinical research: applications to practice. Philadelphia, PA: Davis Company; 2015.
16. Johnson A, Goldfine A. Intrasubject reliability of maximum phonation time. *J Voice*. 2016;30:775.e1–775.e4.
17. Reich A, Mason J, Polen S. Task administration variables affecting phonation-time measures in third-grade girls with normal voice quality. *LSHSS*. 1986;17:262–9.
18. Speyer R, Bogaardt HCA, Passos VL, Roodenburg NPHD, Zumach A, Heijnen MAM, et al. Maximum phonation time: variability and reliability. *J Voice*. 2010; 24:281–4.
19. Ball LJ, Willis A, Beukelman DR, Pattee GL. A protocol for identification of early bulbar signs in amyotrophic lateral sclerosis. *J Neurol Sci*. 2001;191:43–53.
20. Maryn Y, De Bodt M, Roy N. The Acoustic Voice Quality Index: toward improved treatment outcomes assessment in voice disorders. *J Commun Disord*. 2010;43:161–74.