

An Automated Listener-Effort Measure for ALS Generalizes across Diverse Platforms and Protocols

Esteban G. Roitberg

Peter Cohen Foundation dba Everything ALS

Marcos A. Trevisan

Universidad de Buenos Aires

Julian Peller

`julian@everythingals.org`

Peter Cohen Foundation dba Everything ALS

Federico Seveler

Peter Cohen Foundation dba Everything ALS

Felipe Aguirre

Peter Cohen Foundation dba Everything ALS

Diego E. Shalom

Universidad de Buenos Aires

Alan Taitz

SRI International

Gaston Bujia

Peter Cohen Foundation dba Everything ALS

Joel Schwartz

Bristol-Myers Squibb (United States)

Jason Osik

Bristol-Myers Squibb (United States)

Ryan Shewcraft

Bristol-Myers Squibb (United States)

Anahita Kyani

Johnson & Johnson Innovative Medicine

Guofa Shou

Eli Lilly and Company

Nader Naghavi

Regeneron Pharmaceuticals

Henk Schuring

Prilenia Therapeutics B.V.

Vinni Bijanki

Massachusetts General Brigham

Lindsay Heyd

Massachusetts General Brigham

Brittney Harkey

Massachusetts General Brigham

Taylor K. Stimpson

Massachusetts General Brigham

Marianne Chase

Massachusetts General Brigham

Hong Yu

Massachusetts General Brigham

Alexander V. Sherman

Massachusetts General Brigham

Jeremy Shefner

Barrow Neurological Institute

Sabrina Paganoni

Massachusetts General Brigham

Merit E. Cudkowicz

Massachusetts General Brigham

Eric Macklin

Massachusetts General Brigham

Jordan R. Green

Massachusetts General Hospital

Oren Levy

Regeneron Pharmaceuticals

Ernest Fraenkel

Massachusetts Institute of Technology

Lyle W. Ostrow

Temple University

Indu Navar Bingham

Peter Cohen Foundation dba Everything ALS

James D. Berry

Massachusetts General Brigham

Article

Keywords: Amyotrophic Lateral Sclerosis, Listener Effort, Dysarthria, Machine Learning, External Evaluation

Posted Date: July 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9849001/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: Competing interest reported. The authors declare the following competing interests. This work was conducted while ER and AT were paid consultants working at EverythingALS. ER is currently an employee of ITBA, and AT is currently an employee of SRI International. JP, MT, FS, GB, FA and DS are paid consultants to EverythingALS. JO and JSc are current employees of Bristol Myers Squibb. This work was conducted while RS was an employee of Bristol Myers Squibb. NN and OL are current employees and holds stock options in Regeneron Pharmaceuticals, Inc. OL serves as an unpaid scientific advisor to EverythingALS. HS has received personal compensation as an employee of Prilenia Therapeutics B.V. and holds stock options in the company. AK is an employee of J&J Innovative Medicine and holds stock options in the company. GS is an employee of Eli Lilly & Co. and holds company stock. AS has received grants and contracts for clinical research projects sponsored by the FDA, NIH/NIA, NIH/NINDS, The ALS Association, and ALS Finding a Cure Foundation, as well as study support from Shionogi, Inc., Biogen, and Amylyx. JSh is a consultant for Acurastem, Nura Bio, and have received research funds from the Healey Center. MCu is a consultant for Otsuka, Qoralis, Regeneron, Immunity Pharma, on board of directors of Praxis Precision Medicine and have received institutional grants from Prilenia for Healey ALS Platform Trial. SP reports research grants from Amylyx Therapeutics, Revalesio, Eledon, Alector, UCB Pharma, Biohaven, Clene Nanomedicine, Prilenia Therapeutics, Seelos, Calico, Denali, Neurizon, NIH, DoD, the Muscular Dystrophy Association and reports consulting fees from Amylyx, Arrowhead, Biogen, BMS, Clene, Cytokinetics, Eikonizo, J&J, Merck, PharmAust, Prilenia, and Sola. She has been a paid educational speaker for Medscape, PeerView, and i3Health. EF and LO serve as unpaid scientific advisors to EverythingALS. JB has received research support from Rapa Therapeutics, MT Pharma of America, ProJenX, Novartis, MDA, ALS Finding a Cure, the ALS Association, the Association for Frontotemporal Dementia and NINDS. JB has served as a consultant to Alexion Pharmaceuticals, Amylyx Pharmaceuticals, Biogen, MTPA, Regeneron Pharmaceuticals, Roon, Sanofi and Trace Neuroscience. JB also serves as an unpaid scientific advisor to EverythingALS. INB is the founder of EverythingALS

An Automated Listener-Effort Measure for ALS

Generalizes Across Diverse Platforms and Protocols

Esteban G. Roitberg, PhD^{1,2,3}, Marcos A. Trevisan, PhD^{4,5}, Julian Peller, MSCS¹, Federico Sevliver, PhD¹, Felipe Aguirre¹, Diego E. Shalom, PhD^{4,5}, Alan Taitz, PhD⁶, Gaston Bujia, MSc¹, Joel Schwartz, PhD⁷, Jason Osik, PhD⁷, Ryan Shewcraft, PhD⁷, Anahita Kyani, PhD⁸, Guofa Shou, PhD⁹, Nader Naghavi, PhD¹⁰, Henk Schuring, PharmD¹¹, Vinny Bijanki¹², Lindsay Heyd¹², Brittney Harkey¹², Taylor K. Stimpson¹², Marianne Chase¹², Hong Yu, MS¹², Alexander V. Sherman, MS^{12, 14}, Jeremy Shefner, MD, PhD¹³, Sabrina Paganoni, MD, PhD^{12, 14}, Merit E. Cudkiewicz, MD, MSc^{12, 14}, Eric Macklin, PhD¹², Jordan R. Green, PhD¹⁵, Oren Levy, MD, PhD¹⁰, Ernest Fraenkel, PhD¹⁶, Lyle W. Ostrow, MD, PhD¹⁷, Indu Navar Bingham, MSCS¹, James D. Berry, MD, MPH^{12,14}

¹Peter Cohen Foundation dba Everything ALS, Seattle, Washington and Los Altos, California, USA; ²Instituto Tecnológico de Buenos Aires (ITBA), Buenos Aires, Argentina; ³Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, Buenos Aires, Argentina; ⁴Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina; ⁵Instituto de Física Interdisciplinaria y Aplicada, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires – CONICET, Buenos Aires, Argentina; ⁶SRI International, Menlo Park, CA, USA (this work was performed working at ¹); ⁷Bristol Myers Squibb, Cambridge, MA, USA; ⁸Johnson & Johnson Innovative Medicine, La Jolla, CA, USA; ⁹Eli Lilly and Company, Boston, MA, United States; ¹⁰Regeneron Pharmaceuticals, Tarrytown, NY, United States; ¹¹Prilenia Therapeutics B.V., Naarden, the Netherlands; ¹²Sean M. Healey & AMG Center for ALS, Massachusetts General Brigham; ¹³Barrows Neurological Institute; ¹⁴Harvard Medical School, Boston, MA, USA; ¹⁵MGH Institute of Health Professions, Charlestown Navy Yard, Boston, MA, USA; ¹⁶Massachusetts Institute of Technology, Department of Biological Engineering, Cambridge, MA, USA; ¹⁷Lewis Katz School of Medicine at Temple University, Department of Neurology, Philadelphia, PA, USA;

Introduction. Monitoring ALS progression through speech is challenging due to the complex impact of dysarthria across multiple subsystems. Listener Effort (LE), assessed by Speech-Language Pathologists (SLPs), condenses this complexity into a single, clinically meaningful measure of dysarthria severity. Automating LE prediction could enable scalable monitoring and practical clinical applications.

Methods. We developed a machine learning model that uses speaking rate and Whisper confidence, a transcription confidence score from an Automatic Speech Recognition model, to predict SLP-rated LE from remotely collected speech recordings. Speech samples derived from three ALS study datasets - Austen Study (1,656 recordings from 105 participants), HEALEY ALS Platform Trial (4,050 from 266), and Radcliff Study (1,488 from 57) - were annotated with LE ratings by three SLPs. The original model,

trained on Austen, was externally validated on the HEALEY ALS Platform Trial and Radcliff, then updated with data from all three datasets.

Results. The model accurately predicted LE in the external datasets (HEALEY ALS Platform Trial: $R^2 = 0.81 \pm 0.02$; Radcliff: $R^2 = 0.82 \pm 0.03$). Longitudinal analyses showed strong agreement between SLP and model predictions, with high correlations in individual slopes (HEALEY ALS Platform Trial: $CCC = 0.80$; Radcliff: $CCC = 0.93$).

Discussion. We show that machine learning-based LE prediction provides a scalable, reliable, and efficient approach for remote monitoring of ALS using speech. Notably, the model performed well across three independently collected datasets that differed in recording platforms, speech protocols, speech materials, and participant characteristics. This consistently robust performance with different speech data and collection methods suggests the potential for reliable deployment across diverse studies and clinical settings. Further, the LE prediction model offers a sensitive endpoint for clinical trials while reducing reliance on labor-intensive manual SLP ratings. Although SLP assessments remain essential for validation, automated models substantially extend the clinical utility of speech-based measures in ALS research and care.

Keywords: Amyotrophic Lateral Sclerosis; Listener Effort; Dysarthria; Machine Learning; External Evaluation.

1 Introduction

Machine learning algorithms, which learn to generalize analyses from a set of training data, are increasingly used in clinical practice, particularly in the monitoring of diseases that require continuous observation of subtle clinical features (Rajkomar et al., 2019; Esteva et al., 2019; Miotto et al., 2016). Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disorder that progressively impairs motor control. Dysarthria is a hallmark feature of the disease (Hardiman et al., 2017; Kiernan et al., 2011; Chio et al., 2016). Monitoring ALS progression through speech is challenging due to the involvement of multiple speech subsystems, the difficulty of automatically extracting reliable voice

markers, and the labor-intensive nature of manual speech assessments (Beukelman and Mirenda, 2011; McGarrigle et al., 2017). One promising metric in ALS progression is Listener Effort (LE), a subjective rating from 0-100 of the effort it takes a listener to understand dysarthric speech (McGarrigle et al., 2017; Yorkston et al., 1996; Merler et al., 2025; Bingham et al., 2025). LE was originally developed to be assessed by trained Speech and Language Pathologists (SLP-rated LE). It is inherently clinically meaningful – given that the purpose of speech is to be understood. SLP-rated LE strongly correlates with the ALSFRS-R speech subscore, but has a finer granularity and is more sensitive to early changes in speech (Merler et al., 2025; Bingham et al., 2025). As such, it has the potential to serve as an outcome measure for tracking disease progression (McGarrigle et al., 2017).

Since automated machine learning LE assessment (ML-predicted LE) does not require manual rating of speech recordings by (often multiple) highly trained SLPs, it has the potential for easier scalability and cost effectiveness than SLP-rated LE, allowing for a broader application in ALS trials and different clinical settings (Merler et al., 2025; Bingham et al., 2025). Machine learning models must be appropriately developed and tested, since once they are deployed, they are expected to generalize to new data. Such model training is typically accomplished using a single dataset and holding out a portion of the data for subsequent validation, bootstrapping (repeatedly re-sampling the original dataset with replacement), or cross validation techniques. All these strategies involve training on a dataset while withholding a portion of the data, which is set aside to test the model after training. These processes are broadly referred to as *internal validation* (Cabitza et al., 2021).

Internal validation may not be sufficient to demonstrate that models will perform well when applied to new datasets and data collection methods. *External validation* or *external evaluation* (Bleeker et al., 2003; Cabitza et al., 2021; Collins et al., 2024a) involves validating models with entirely different data and can be particularly important with the high dimensionality and relatively small datasets encountered in healthcare applications (Vabalas et al., 2019).

Harmonized guidelines support the robust development and validation of machine learning models (Wynants et al., 2020; Collins et al., 2024a). In particular, the *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + AI*, (TRIPOD+AI, Collins et al. (2024b)) includes reporting recommendations on study design, data sources, predictors, outcomes, analytical methods, performance measures, and limitations. In the current study, we employed the TRIPOD+AI

framework to evaluate the robustness of the machine learning Listener Effort model published in Bingham et al. (2025), for predicting SLP-rated LE scores and monitoring ALS progression across datasets. We analyzed model performance on speech recordings from people living with ALS (pALS) collected by three independent studies: 1) the Austen Study (Merler et al., 2025), the original training dataset; 2) the HEALEY ALS Platform Trial (Regimen D, plus placebo data from Regimens A, B, and C), hereafter referred to as *HEALEY* (Harkey et al., 2025; Paganoni et al., 2022, 2024, 2025; Writing Committee for the HEALEY ALS Platform Trial 2025a, b, c), and 3) the Radcliff Study (Peller et al., 2025).

The three studies were designed and conducted independently and vary in cohort size, session frequency, and study duration. Perhaps more importantly, they also differ in recording platforms and speech protocols. The speech samples themselves vary in length, content, number of utterances, and lexical composition (see Methods). These characteristics introduce multiple sources of variability commonly encountered in real-world applications. Crucially, these datasets also include participants with varying distributions of dysarthria severity. Together, these datasets provide a particularly challenging benchmark for evaluating listener effort generalizability. Strong performance across all three studies would indicate that the measure remains robust despite substantial differences in study design, recording conditions, speech material, and disease severity.

Characteristic	Austen	HEALEY	Radcliff
Participants, n	105	266	57
Time in study, months (median [IQR])	5.5 [1.1–11.1]	5.1 [3.8–5.6]	6.0 [1.1–10.8]
BQC sessions, n	584	1519	521
AQC sessions, n	552	1350	496
AQC sessions per participant (mean $\pm$ SD)	5.26 $\pm$ 3.16	5.08 $\pm$ 2.23	8.70 $\pm$ 6.84
BQC speech samples, n	1752	4557	1563
AQC speech samples, n	1656	4050	1488

Table 1: Participation. Total number of patients per study. Multiple recording sessions were held per patient and several speech samples were recorded per session. Counts before and after quality control filters (BQC and AQC), as described in Methods section, are shown.

A linear regression model, trained on two acoustic features from the Austen dataset, was externally validated on HEALEY and Radcliff. The model was then updated using all three datasets and re-evaluated. To assess clinical utility, longitudinal analyses were performed on both SLP-rated LE and ML-predicted LE scores, comparing their agreement in estimating rates of progression. The primary objective was to determine the generalizability of these models across ALS populations, using both internal and external validation, and to benchmark their performance against SLP-based assessments. Ultimately, we investigate whether ML-predicted LE scores can serve as a reliable and scalable tool for ALS monitoring while maintaining close alignment with expert clinical evaluations of speech progression.

2 Results

2.1 Participant Characteristics and Dataset Overview

The number of participants and analyzed recordings are summarized in Table 1. Additionally, Figure 1 illustrates the timespan and all recording sessions for each participant, distinguishing between all sessions, those scored by SLPs, and those discarded by the quality control filters detailed in the Methods section.

The demographic data of the participants for the three studies is summarized in Table 2. Most of the participants were older adults, white, and non-Hispanic. Austen and Radcliff studies were gender balanced. At baseline, participants in the HEALEY study exhibited more advanced ALS-related symptoms, as indicated by ALSFRS-R scores, subscores, speaking rates and the time between the onset of symptoms and the enrollment in the study (Cedarbaum et al., 1999).

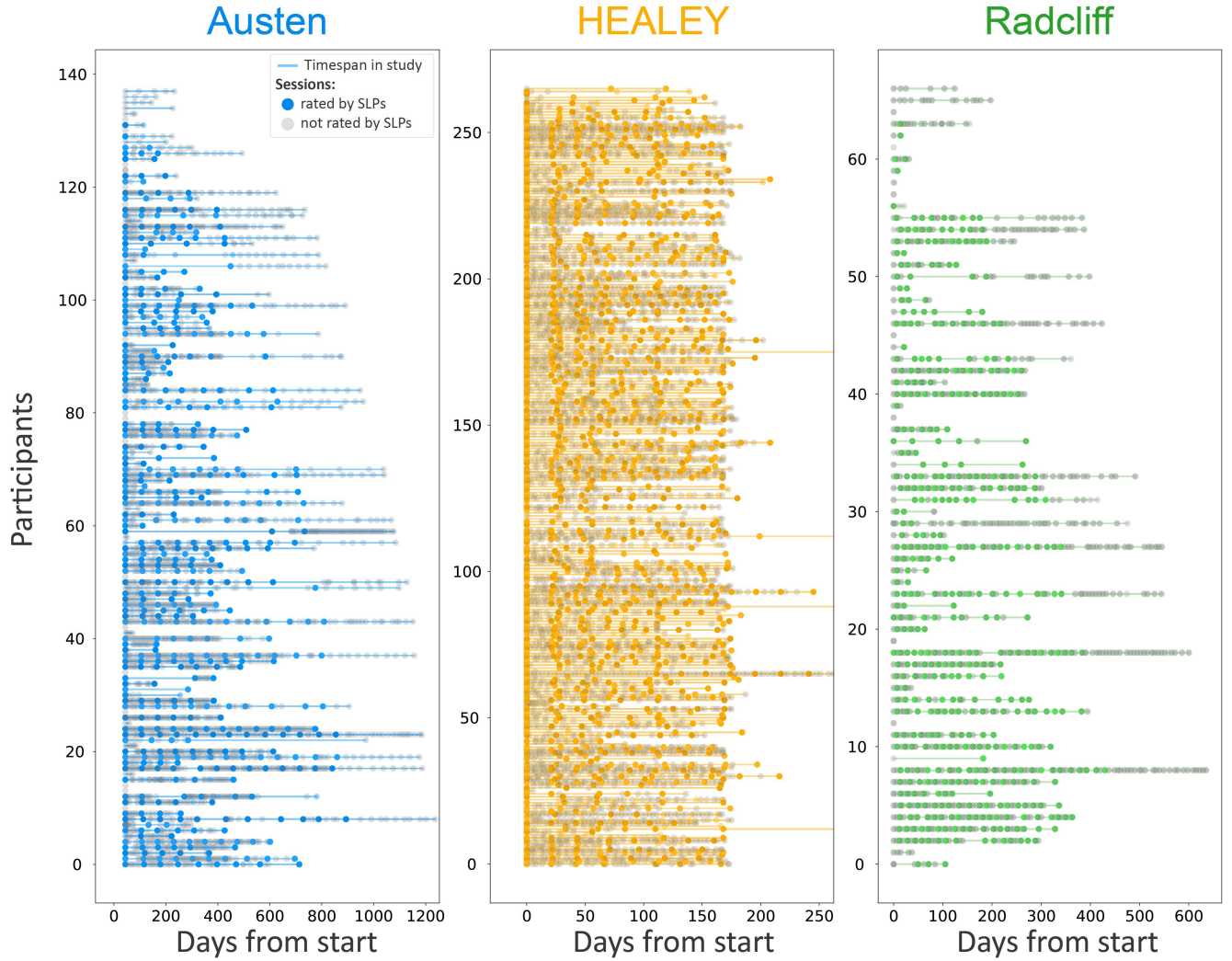


Figure 1: Participant Flow. Participation timelines for the three studies. Each horizontal line represents a participant, and dots indicate recording sessions (darker shades of the same colors indicate sessions rated by multiple Speech Language Pathologists (SLP); unrated sessions are shown with light gray). The datasets analyzed in this work differed in recording platforms, protocols, speech content, cohort size (105, 266 and 57 pALS), speech session frequency (median intervals of 67, 24 and 17 days between rated sessions), and study duration (mean of 232, 141 and 197 days).

Characteristic	Austen	HEALEY	Radcliff
Participants, <i>n</i>	105	266	57
Age, years (median [IQR])	67 [39–83]	60 [52–66]	66 [54–70]
Sex, <i>n</i>			
Female	50	94	23
Male	55	172	29
Unknown / not reported	0	0	5
Ethnicity, <i>n</i>			
Non-Hispanic	98	250	49
Hispanic	5	13	2
Race, <i>n</i>			
White	94	242	48
Asian	5	5	3
Other	4	14	1
ALSFRS-R total at enrollment (mean $\pm$ SD)	42.4 $\pm$ 7.6	35.0 $\pm$ 6.8	40.6 $\pm$ 5.1
ALSFRS-R total at final assessment (mean $\pm$ SD)	40.8 $\pm$ 9.4	30.2 $\pm$ 8.4	38.7 $\pm$ 5.6
ALSFRS-R Q1 (Speech) at enrollment (mean $\pm$ SD)	3.66 $\pm$ 0.69	3.09 $\pm$ 1.02	3.41 $\pm$ 0.73
ALSFRS-R Q1 (Speech) at final assessment (mean $\pm$ SD)	3.59 $\pm$ 0.82	2.68 $\pm$ 1.23	3.27 $\pm$ 0.83
ALSFRS-R Bulbar score at enrollment (mean $\pm$ SD)	11.2 $\pm$ 1.5	10.0 $\pm$ 2.2	10.4 $\pm$ 1.9
ALSFRS-R Bulbar score at final assessment (mean $\pm$ SD)	10.9 $\pm$ 0.9	9.0 $\pm$ 3.0	10.0 $\pm$ 2.1
ALSFRS-R slope at enrollment (points/month) (mean $\pm$ SD)	0.44 $\pm$ 0.62	0.69 $\pm$ 0.45	0.27 $\pm$ 0.23
Speaking rate, words/s (mean $\pm$ SD)	2.62 $\pm$ 0.76	2.49 $\pm$ 0.84	2.93 $\pm$ 0.87
Time from symptom onset to diagnosis, years (mean $\pm$ SD)	1.4 $\pm$ 1.5	0.9 $\pm$ 0.5	1.1 $\pm$ 3.5
Time from symptom onset to enrollment, years (mean $\pm$ SD)	4.1 $\pm$ 3.8	1.8 $\pm$ 0.7	2.2 $\pm$ 4.9
Reported ALS inheritance, <i>n</i>			
Familial	40	33	7
Sporadic	65	226	27
Onset location, <i>n</i>			
Bulbar	23	46	10
Non-bulbar	81	220	41

Table 2: Participant demographics. Demographic characteristics are shown per study. Most participants were older adults, White, and non-Hispanic. Austen and Radcliff were gender balanced (five participants did not report sex). Participants in the HEALEY study exhibited more advanced ALS-related symptoms, as indicated by ALSFRS-R scores, subscores, speaking rates and the time between symptom onset and study enrollment.

2.2 Assessment of SLP Scoring Consistency and Objectivity

The ICCs for inter- and intra-rater reliability (Koo and Li, 2016) exceed 0.9, indicating a high level of agreement in both (Table 3). These results support the robustness of SLP scores as a reliable ground truth for model development.

	Level	Study	ICC
Intra-rater	Task	Austen	0.91 [0.89; 0.92]
		HEALEY	0.93 [0.91; 0.95]
		Radcliff	0.91 [0.88; 0.95]
	Session	Austen	0.91 [0.89; 0.93]
		HEALEY	0.97 [0.96; 0.98]
		Radcliff	0.97 [0.95; 0.98]
Inter-rater	Task	Austen	0.90 [0.90; 0.90; 0.90]
		HEALEY	0.92 [0.92; 0.92; 0.92]
		Radcliff	0.93 [0.93; 0.93; 0.93]
	Session	Austen	0.95 [0.95; 0.95; 0.95]
		HEALEY	0.96 [0.96; 0.96; 0.96]
		Radcliff	0.97 [0.97; 0.97; 0.97]

Table 3. Intra- and inter-rater reliability of SLP scores. SLP score reliability was assessed using intra- and inter-rater ICCs. **Intra-rater ICC** was calculated by comparing repeated ratings of the same recordings by the same SLP, corresponding to 20% of the sentences. The task level shows ICCs across all repeated ratings, whereas the session level shows ICCs across session-averaged repeated ratings. Brackets report 95% confidence intervals. **Inter-rater ICC** was calculated by comparing pairwise ratings from the three SLPs on the same recordings. The mean across SLP pairs is shown, and brackets report each pairwise value. The task level shows ICCs across all rated pairs, whereas the session level shows ICCs across session-averaged pairs. Each session included three tasks. For both intra- and inter-rater ICCs, single-task variability was smoothed at the session level, resulting in higher reliability.

2.3 Model Features and Listener Effort Distributions

The machine learning model developed in (Navar Bingham et al., 2025) relied on two features extracted from each speech sample: speaking rate and Whisper confidence, the latter reflecting Whisper's confidence in the automatic transcription (see Methods). The distributions of the input and output variables across the three studies are shown in Figure 2. Participants in the HEALEY study exhibit lower Whisper confidence, slower speaking rates, and higher ML-predicted LE scores – consistent with the more advanced symptoms of HEALEY participants demonstrated by ALSFRS-R scores at time of enrollment (see Table 2) compared to those enrolled in the Austen and Radcliff studies.

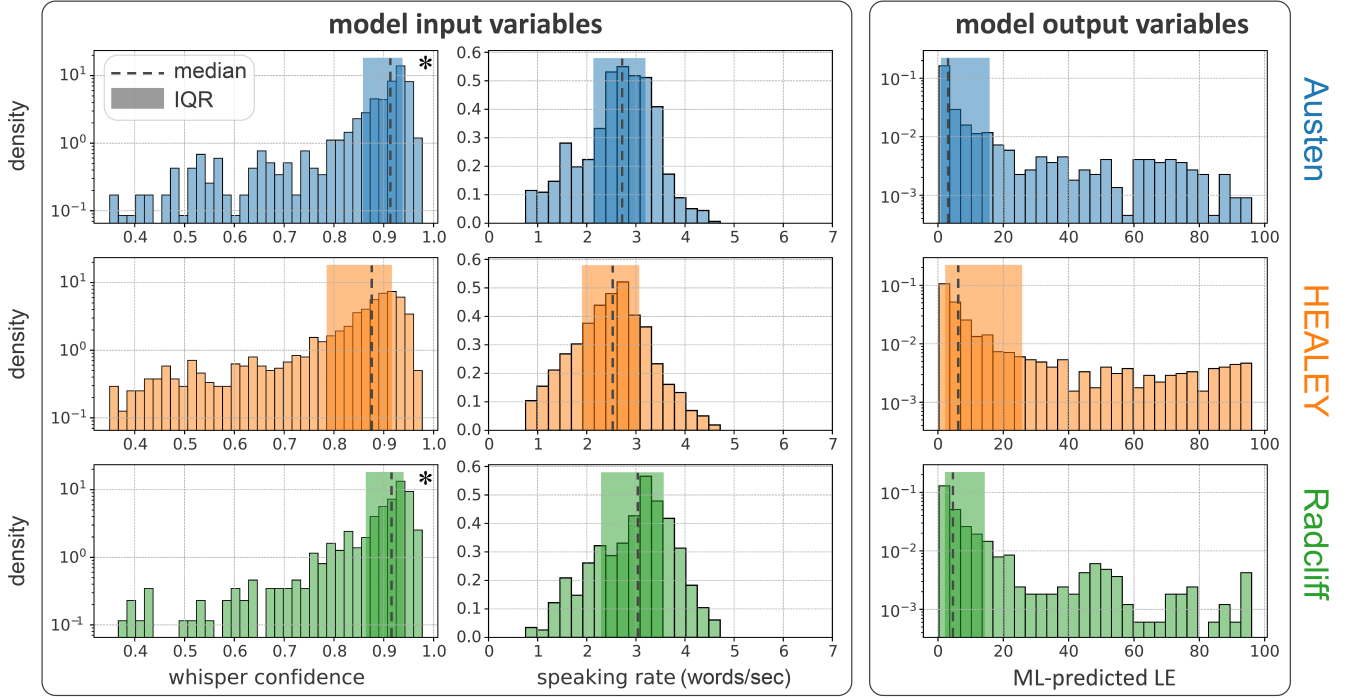


Figure 2: Feature Distributions. Distributions of input and output variables used in the LE model across the three studies. From left to right, panels show Whisper confidence (log scale), speaking rate (linear scale), and ML-predicted LE (log scale). From top to bottom, panels correspond to the Austen, HEALEY, and Radcliff studies. Compared to the HEALEY, both Austen and Radcliff Whisper confidence distributions are higher, consistent with differences in ALSFRS-R score distributions (see Table 2). Asterisks indicate that a paired t-test did not detect a significant difference between distributions.

2.4 External Evaluation

Results comparing ML-predicted LE, trained on the Austen dataset, with the SLP-rated LE are shown in Figure 3 (top row). Because the transcription-accuracy filter removes most severe-dysarthria recordings (10% of participants with SLP-rated LE between 80 and 100 did not receive a ML-predicted LE score), these values reflect performance on a more transcribable subset and likely overstate accuracy on unselected clinical speech. Overall, the model demonstrated good performance, with an R^2 above 0.8 and a Root Mean Square Error ($RMSE$) below 10. The highest R^2 scores were observed in the Austen dataset, as expected, since it was the training population. The poorest performance, based on both R^2 and $RMSE$, was found in the HEALEY dataset. This was a clinical trial with a focus on enrolling people with faster progression rates; speaking rate and Whisper confidence were both significantly lower and ML-predicted LE was significantly greater (Figure 2). Standard deviations for both R^2 and $RMSE$ were estimated using cross-validation for the Austen dataset (used in model development) and bootstrapping for the HEALEY and Radcliff datasets (used for external evaluation), as suggested in (Collins et al., 2024a). Since nearly all deviations are below 10% of the corresponding value, we conclude that the dataset sizes were adequate, and that the model predictions effectively mirror those obtained from SLP ratings, both at the individual and population levels.

2.5 Model Update

The model trained in the original Austen dataset was then updated using a cross-validation technique with random folds, shuffling data from all three datasets to improve generalizability (see Methods). Results from the updated model are shown in Figure 3 (bottom row). This approach is expected to reduce performance on the original dataset, due to decreased specificity, while improving performance on new, unseen patient cohorts by increasing generality.

In fact, the updated model shows a trade-off between performance and generalizability. While cross-validation results indicate slightly lower R^2 values compared to the development model, $RMSE$ is improved in some datasets (e.g., HEALEY), suggesting modest gains in prediction accuracy. These differences highlight that model updating does not universally improve fit to all datasets but rather balances performance across populations. Importantly, the updated model is expected to provide more

robust and generalizable predictions when applied to new patient cohorts, making it the preferred choice for external applications where unseen variability is inevitable (Van Calster et al., 2023).

While further external validation in additional cohorts may always be informative, validation of prediction models is inherently an ongoing process rather than a finite step, and the current evaluation across multiple independent datasets provides a sufficiently robust assessment of its generalizability for the intended use.

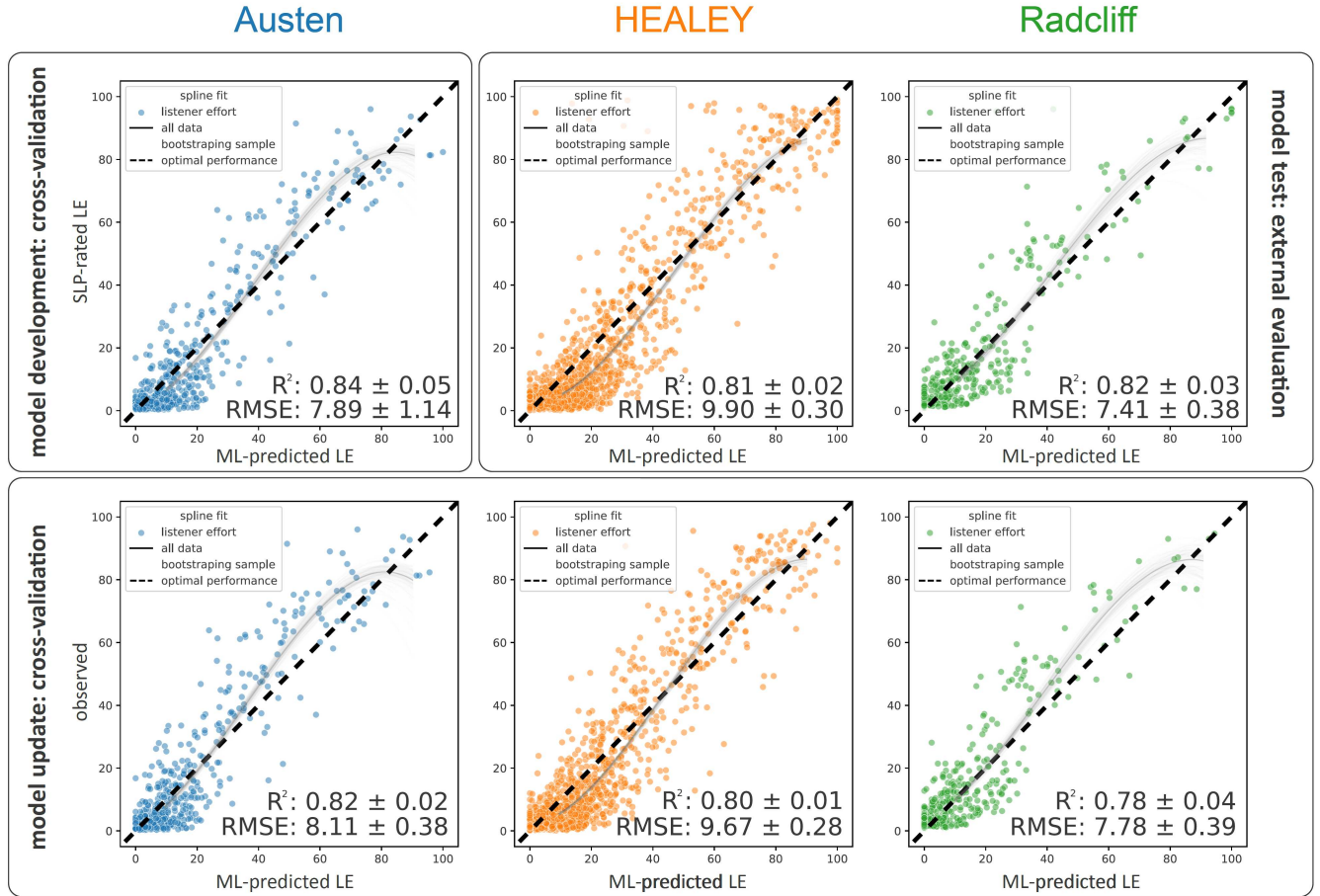


Figure 3: SLP-rated vs. ML-predicted Listener Effort across Studies and Evaluation Stages. Each panel shows a scatter plot comparing SLPs and model values of LE for the Austen (blue), HEALEY (orange), and Radcliff (green) datasets. In the top row, model development and cross-validation were performed exclusively on the Austen dataset (left). The HEALEY and Radcliff panels (center and right) correspond to external evaluations using the model trained on Austen data. The bottom row shows model performance after updating with all datasets. Each dot represents a single task for one participant. Gray curves represent spline fits to bootstrapped resamples. The dashed diagonal line indicates the optimal (identity) fit. Model performance is quantified by R² and root mean squared error (RMSE), reported in the lower-right corner of each panel. Both values in the Austen cross validation panel (top left) are computed from cross-validation folds, while all other metrics are derived from bootstrap resampling.

2.6 Data Size, Bootstrap, and Spline Curves

To quantify the expected deviations in performance measurements, particularly their dependence on dataset size, we assessed whether our three datasets were sufficiently large by applying bootstrap resampling and spline interpolation to both the external evaluation and model update stages, as recommended in Riley et al. (2024). Bootstrap analysis provided standard deviations for both R^2 and $RMSE$ (shown in Figure 3), as well as spline curves for the original datasets and the bootstrap samples.

The distribution of bootstrap-derived spline curves offers insights into the expected variability attributable to dataset size. Deviations were negligible across most of the LE range (0–90). In contrast, values in the 90–100 range exhibited greater dispersion, reflecting reduced robustness of model performance. This was anticipated given the scarcity of participants in this range, as well as the bias introduced by QC filters, which excluded recordings with poor transcriptions.

2.7 Longitudinal Analysis

Once the agreement between SLP-rated and ML-predicted LE was established - particularly when using the model trained on Austen dataset - longitudinal analyses were performed on the external evaluations to assess clinical utility.

We fitted linear mixed-effects models (LMM) for each dataset, comparing SLP-rated and ML-predicted LE along with the population linear regression in individual participant slopes. Results are shown in Figure 4. Population progression slopes were similar in magnitude and direction for the two measures, and paired t-tests did not detect a significant difference in either dataset. We note that a non-significant difference does not by itself establish equivalence; rather, these results indicate that SLP-rated and ML-predicted LE yield consistent estimates of the direction and approximate rate of progression in these cohorts. Notably, slopes from Radcliff are smaller than those from HEALEY for both SLP-rated and ML-predicted values, reflecting differences in the study populations: HEALEY, as a clinical trial, tends to enroll faster progressors, whereas Radcliff targeted earlier-stage pALS (see Methods).

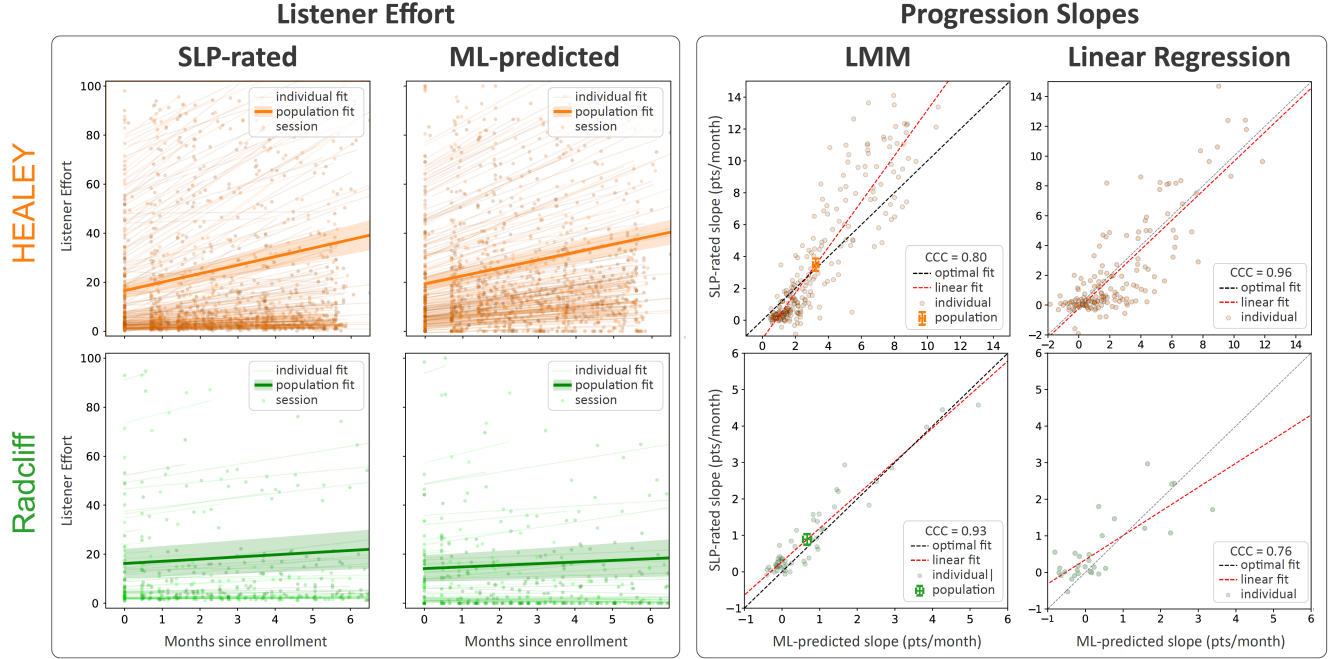


Figure 4: Longitudinal analysis on SLP-rated LE vs. ML-predicted LE. Comparison of LE progression using the Austen-trained model on the HEALEY (top panels) and Radcliff study (bottom panels). Left panels show LE scores, both SLP-rated (left column) and ML-predicted (center-left column) over time. Each dot represents one recording session mean score. Right panels show the individual slopes derived from the SLP-rated and ML-predicted LE values, using both a linear mixed-effects model (LMM, center-right column) and linear regression (right column). Using an LMM, high Concordance Correlation Coefficients were observed ($CCC = 0.80$ and $CCC = 0.93$), indicating that ML-predicted LE scores mirror SLP-rated LE. Intercepts (not shown) also align closely. Population slope estimates are also consistent. For both slopes and intercepts, we applied a paired t-test to assess similarity between SLP-rated and ML-predicted LE values and we found no significant differences ($p > 0.1$). These results suggest that conclusions drawn from the model predictions mirror those obtained from SLP ratings, both at the individual and population levels. Linear regression provided similar results, maintaining strong agreement ($CCC = 0.96$ and $CCC = 0.76$).

When analyzing participant-specific slopes in both HEALEY and Radcliff, the slopes obtained from the LMM fitted to SLP-rated LE and those from the LMM fitted to ML-predicted LE are highly correlated with $CCC = 0.80$ and $CCC = 0.93$, respectively. HEALEY CCC was slightly lower, we hypothesize that this may be due to the heterogeneity of the HEALEY population, which includes both fast and slow progressors, while the LMM estimates an overall population-level progression, thereby overestimating slow progressors and underestimating fast progressors (HEALEY slopes $median = 0.59 \frac{pts}{month}$, $IQR = 2.60 \frac{pts}{month}$; Radcliff slopes $median = 0.15 \frac{pts}{month}$, $IQR = 1.03 \frac{pts}{month}$). To assess this issue, we conducted a simple linear regression for each individual, which revealed a wider range of progression

slopes and a higher correlation in HEALEY. Moreover, as Radcliff study consists of a more homogeneous population, the LMM *CCC* is consistently greater than the linear regression *CCC* (0.96 and 0.76 respectively).

Overall, these longitudinal analyses demonstrate that ML-predicted LE scores closely track ground-truth ratings over time, both at the population and individual levels. The consistency of results across datasets supports the utility of ML-predicted LE as a complement for SLP-rated LE in monitoring disease progression.

3 Discussion

This study shows that a simple and interpretable machine learning model, based only on speaking rate and Whisper confidence, can reliably estimate Listener Effort in individuals with ALS. The model demonstrated strong external validation across independent cohorts that differed in demographics, disease characteristics, speech protocols and data collection platforms (Modality.ai in Austen, Aural Analytics in HEALEY and Radcliff) supporting its robustness and clinical applicability. These results reinforce the potential of ML-predicted Listener Effort as a multicenter endpoint in ALS research.

From a methodological perspective, the findings emphasize the importance of rigorous external validation for machine learning models intended for clinical use. Longitudinal analyses showed high agreement between SLP-rated and ML-predicted Listener Effort, both at the group level and in individual trajectories, highlighting the model's value for tracking disease progression over time. We found strong correlation of progression slopes. This indicates that both SLP-rated LE and the ML-predicted LE consistently identify whether participants are progressing. However, they slightly differ in estimating the magnitude of that progression. The ML model tends to produce finer, slightly higher estimates for individuals with little to no dysarthria, and more moderate values for those with severe dysarthria, compared to SLP-rated LE. This effect is reflected in the progression slopes, which are overestimated for slow progressors with low LE values. If these finer gradations in ML-predicted LE reflect real differences in speech impairment, they could improve the ability to detect dysarthria onset at early stages. Future comparative analyses between individuals with early ALS and healthy controls could help clarify this possibility.

Regarding cohort heterogeneity, the LMM’s assumption of a single global trajectory leads to overestimation in slow progressors and underestimation in fast ones. Overall, these results suggest that the LMM has limited precision in estimating individual progression rates, and caution is warranted when applying LMMs to heterogeneous populations.

From a clinical perspective, ML-predicted LE has several important advantages. It enables scalable and remote monitoring of dysarthria progression in ALS. Because LE reflects the everyday communication challenges faced by patients, the model provides sensitive, patient-centered endpoints that extend beyond traditional clinical scales such as the ALSFRS-R. Furthermore, ML-predicted LE offers reliable quantitative markers for clinical trials, with the potential to capture subtle therapeutic effects that may otherwise go undetected.

Nonetheless, some limitations must be acknowledged. The model was trained and validated primarily on controlled speech tasks, which may not fully capture the variability of spontaneous or naturalistic speech in daily life. In addition, extremely severe dysarthria was underrepresented due to transcription-based quality filters (10% of participants with SLP-rated LE between 80 and 100 did not receive a ML-predicted LE score). Finally, the model has not yet been tested in non-English speaking populations.

Future work should expand training and evaluation datasets to include spontaneous speech. In addition, recent improvements in transcription algorithms may help address the underrepresentation of recordings with extreme dysarthria. These are extreme cases in which the ML-predicted LE pipeline is affected by technical and contingent artifacts related to transcription, while clinical assessment is relatively straightforward. Moreover, LE scores in the 80 to 100 range correspond to final, rather than transient, progression scores and are therefore less informative for measuring severity progression.

Although the datasets include speakers from multiple U.S. regions, the overwhelming majority of participants in the analyzed datasets are white native English speakers. As a result, our findings may not fully capture the variability of accents and speech patterns present in more diverse populations. This limitation is particularly relevant given documented disparities in ASR performance across demographic groups. For example, commercial ASR systems have shown higher word error rates for Black speakers (35%) than for white speakers (19%) (Koenecke et al., 2020), and modern ASR models continue to struggle with accented and non-standard dialects such as African American Vernacular English and

Chicano English (Harris et al., 2024). These findings highlight the importance of external validation across diverse speaker populations to identify potential biases and ensure robust generalization.

Integrating ML-predicted LE with other digital markers, such as articulatory, respiratory features or wearable-derived motor endpoints, could provide a more comprehensive assessment of disease progression. Embedding the model into telehealth platforms would enable real-time monitoring, patient self-tracking, and clinical decision support, thereby broadening its utility. In the long term, establishing ML-predicted LE as a scalable, clinically meaningful endpoint has the potential to enhance patient care and accelerate therapeutic development in ALS.

4 Methods

4.1 Speech Recordings

4.1.1 Austen Dataset

Speech data were collected remotely using the Modality.ai platform (<https://modality.ai/>). The primary task analyzed in the present manuscript, *rotating sentences*, involved reading phonetically balanced sentences randomly drawn from three pools (19, 16, and 18 sentences of 11, 13, and 15 words).

Additional tasks included sustained vowel phonation, paragraph reading (Bamboo Passage Yunusova et al. (2011); Barnett et al. (2020)), diadochokinetic (DDK) rate (*puh-tuh-kuh* Duffy et al. (2012)), and picture description. Rotating sentences were selected for LE assessment as they minimize learning effects, control utterance length and predictability, and are established in dysarthria research (Allison et al., 2019). Short sentences also improve efficiency for participants and raters.

4.1.2 HEALEY and Radcliff Datasets

Data were collected using AuralAnalytics (<https://auralanalytics.com/>). Three rotating sentence pools (from five available, each with 50 sentences) were used. Other tasks included sustained vowel phonation, fixed reading (five sentences), DDK (*buttercup*), and picture description.

Session intervals ranged from 2–60 days (median = 9 days). The first session per participant was included, followed by sessions at least 21 days apart.

4.2 Ethics Oversight

All aspects of the design and conduct of the Austen Study and the Radcliff Study were conducted under the approval of the Western IRB. All aspects of the design and conduct of the HEALEY Study were done under the approval of the Massachusetts General Brigham IRB. Every participant is presented with a written informed consent through a digital study portal and provides written documentation of informed consent prior to undergoing any study procedure. All participants consented to inclusion of their study data (for example, coded identifiers, demographics, ALS history and outcome measures, and speech samples) in a large speech database that can be accessed by researchers under appropriate data use agreements.

4.3 SLP Scoring

Speech recordings from the three datasets were evaluated by four expert speech-language pathologists (SLPs), three per dataset, using the LE metric - a scale from 0 to 100 indicating the cognitive effort required by a professional listener to understand speech affected by dysarthria. Details of the rating protocol are described elsewhere (Bingham et al., 2025; Merler et al., 2025).

SLPs rated speech samples via a secure, custom web platform using Sennheiser 280 Pro headphones with standardized volume calibration. Two training/pilot phases preceded final scoring (Bingham et al., 2025), resulting in the following protocol:

- Audio normalization to maximum amplitude.
- Volume calibration before rating.
- Platform orientation (login, navigation, rating workflow) with tutorial video.
- Six-sample passive warm-up.
- 20% repeated samples per rater for intra-rater reliability.
- Randomized sample order without consecutive repeats.
- Option to revise previous ratings.

Four independent SLPs participated, with three rating each dataset to assess SLP to SLP variability. Three SLPs each rated one session from 10 subjects (three sentences per subject: 11, 13, and 15 words). With repeats, each rater assessed 36 samples.

SLP-rated LE was scored using a digital Visual Analogue Scale (0 = easily understandable; 100=unintelligible even with full effort), with the slider defaulting to 50. Raters could add free-text notes (e.g., accent, noise), later used in quality filtering. Session-level SLP-rated LE was the mean of nine ratings (three sentences $\times$ three raters).

Before developing the model, we assessed the reliability of the SLP scores. These scores, ranging from 0 to 100, represent the perceived cognitive effort required by a professional listener to understand speech affected by dysarthria. A score of 0 indicates no effort is required, while a score of 100 signifies that all cognitive resources are being deployed in an attempt to recover what is being said. Importantly, listener effort provides a clinical anchor that ensures the model captures change that is meaningful to real-world communication, not merely acoustic variation that may be perceptually irrelevant. These scores serve as ground truth for training machine learning models, and to ensure their suitability for that role, we assessed both consistency and objectivity across all recordings.

Intra-rater reliability is a measure of consistency that refers to the extent to which a given SLP assigns the same score to the same recording upon repeated assessments. To evaluate this, we randomly selected 20% of the recordings (1,438 recordings) and re-presented them to the same SLPs, shuffled within the session to prevent recognition or recall of the original scores. We then quantified the similarity between repeated scores using the Intraclass Correlation Coefficient (ICC) (Koo and Li, 2016), which measures statistical agreement. The ICC ranges from 0 (no agreement) to 1 (perfect agreement).

Inter-rater reliability, a measure of objectivity, reflects the extent to which different SLPs assign the same score to the same recording. We assessed this by applying the same ICC-based analysis to scores from different raters evaluating the same samples. The number of recordings for each dataset is specified in Table 1.

4.4 Quality Controls and Filters

For Austen, Modality.ai applied prespecified QC criteria, excluding incomplete sessions or those with technical issues. HEALEY and Radcliff data were provided unprocessed by the vendor. Regardless of the level of preprocessing by the vendor, three filters were applied to all datasets:

- **SLP notes.** Excluded recordings flagged for quality issues (*noise, volume, error, no speech*).
- **Transcription accuracy.** Retained only recordings with $\geq 20\%$ overlap with the original sentence to ensure reliable feature extraction, acknowledging bias against extreme-dysarthria samples. 59% of sessions with SLP-rated LE between 80 and 100 did not pass this filter, representing 6% of all sessions.
- **Minimum rated recordings.** Excluded sessions with fewer than two usable sentences (i.e., only one sentence with three ratings).

4.5 Listener Effort Predictive Model

Algorithmic computation of LE has been demonstrated as feasible in previous studies (Bingham et al., 2025; Merler et al., 2025). Among the two established approaches, one uses acoustic features together with features derived from automatic transcription software (Whisper, Radford et al. (2023)); these features serve as input to our predictive model.

Our model is a straightforward and interpretable linear regression using two features: Speaking Rate, obtained with Whisper Large-v2 ($SR_{\text{large-v2}}$), and Whisper Confidence, a score ranging from zero to one reflecting the transcription certainty from Whisper Base (WC_{base}). The alternative approach in the literature employs end-to-end models for ML-predicted LE (Merler et al., 2025).

4.5.1 Model Development

The model was first trained using the filtered Austen dataset. The ground truth comprised SLP-rated LE. Mean squared error was used as the loss function. We applied 5-fold cross validation to maximize dataset usage without leaving test recordings outside the training process, following Collins et al. (2024a).

4.5.2 External Evaluation

After development, the model was evaluated on the filtered HEALEY and Radcliff datasets, which were not used in training. This external evaluation assessed generalization performance, as recommended by Collins et al. (2024a).

4.5.3 Model Update

Following evaluation, the model was retrained using the combined filtered data from all three datasets. We again applied 5-fold cross-validation, shuffling recordings from all datasets within each fold. This updated model is expected to perform better on unseen future datasets (Collins et al., 2024a).

4.5.4 Expected Deviations - Bootstrap

To assess expected variability in model performance given each dataset’s sample size, we conducted a bootstrap analysis (Riley et al., 2024). Random samples of the original (SLP-rated LE, ML-predicted LE) pairs were drawn with replacement until the original sample size was reached. R^2 and $RMSE$ were recalculated for each bootstrap replicate, and the process was repeated 100 times to generate distributions for both metrics. Standard deviations of the original R^2 and $RMSE$ values were then computed from these distributions.

4.5.5 Model Calibration - Spline Curves

To investigate model behavior across the LE range, we fitted spline curves using the Univariate Spline function from the `scipy` package in Python (Dierckx, 1995). This function applies cubic polynomial interpolation to the (SLP-rated LE, ML-predicted LE) pairs, producing a smooth curve from 0 to 100. A perfect spline corresponds to the identity line (perfect prediction). In regions where the spline lies above the identity, the model underestimates SLP-rated LE; where it lies below, it overestimates SLP-rated LE.

4.6 Longitudinal Analysis

To further evaluate the agreement between SLP-rated LE and ML-predicted LE, we conducted a longitudinal analysis of both sources of LE (Berry et al., 2019; Stegmann et al., 2021; Johnson et al., 2023). A linear mixed-effects model (LMM) was fitted separately to the SLP-rated LE in the HEALEY dataset and to the ML-predicted LE in HEALEY using the ML model trained on Austen. *Months since enrollment* was included as a fixed effect, with participant-specific random intercepts and random slopes. The fixed effect for *Months since enrollment* (slope) reflects the rate of progression at the population level, while the participant-specific random slopes capture progression at the individual level. If both sources of LE are interchangeable, the fixed-effect parameters (the mean and confidence interval) should be statistically similar, and the individual slopes should be highly correlated.

To assess the similarity of individual slope distributions, we computed the correlation between each participant’s SLP-rated LE and ML-predicted LE slopes. In addition, we performed a paired t-test to evaluate whether the two distributions differ significantly. If the two sources of LE are interchangeable, we would expect a high p -value, indicating no significant difference. Conversely, HEALEY and Radcliff slopes should be different as HEALEY participants have greater values of LE. We applied a two-sample t-test to assess if these cross-dataset distributions are dissimilar, expecting a low p -value. We found no significant differences in any case (HEALEY slopes: SLP-rated $3.5 \frac{pts}{month}$, ML-predicted $3.2 \frac{pts}{month}$, $t(256) = 0.62, p = 0.53$; HEALEY intercepts: SLP-rated 16 pts , ML-predicted 19 pts , $t(256) = 1.15, p = 0.25$; Radcliff slopes: SLP-rated $0.89 \frac{pts}{month}$, ML-predicted $0.67 \frac{pts}{month}$, $t(55) = 1.62, p = 0.11$; Radcliff intercepts: SLP-rated 16 pts , ML-predicted 14 pts , $t(55) = 0.83, p = 0.41$). Slopes from Radcliff are smaller than those from HEALEY for both SLP-rated and ML-predicted values (two-sample unpaired t-test with $t(181) = 11.5, p < 0.001$ for SLP-rated and $t(86) = 4.08, p < 0.001$ for ML-predicted slopes).

An LMM (Berry et al., 2019; Stegmann et al., 2021; Johnson et al., 2023) was fitted separately to the SLP-rated and the ML-predicted LE for both the Radcliff and HEALEY datasets, with *months since enrollment* included as a fixed effect, with participant-specific random intercepts and random slopes. The ML-predicted LE analyzed was generated by the model trained on the Austen dataset; thus, in both Radcliff and HEALEY, we are evaluating external predictions. The objective of these analyses was to

compare the resulting model parameters (individual and population level progression slopes) and evaluate their statistical agreement. If agreement between progression slopes is sufficiently strong, SLP-rated LE and ML-predicted LE should be interchangeable.

Finally, since the LMM estimates a population-level progression slope, we applied a linear fit to each individual's scores versus *Months since enrollment* to compare progression estimates and assess LMM performance in highly heterogeneous populations. This procedure was applied to both SLP-rated LE and ML-predicted LE scores, with the expectation of improved results for the HEALEY dataset, given its greater heterogeneity.

5 Open Science

All the speech recordings and associated de-identified metadata from the Austen study are available to researchers via the Everything ALS Data Portal. Self-reported clinical data, including overall health information, ALS disease history and routine outcome measures, including ALSFRS-RSE and speech recordings (uncompressed .wav format) are hosted in the data portal¹. The repository² contains the data used to generate the figures and tables in this work.

The HEALEY ALS Platform Trial details are accessible through <http://www.clinicaltrials.gov> under identifier NCT04297683. HEALEY ALS Platform Trial placebo data are available upon request³. Placebo data are also available as part of the PRO-ACT database⁴.

6 Acknowledgments

Peter Cohen Foundation would like to thank all the participants in the Austen, HEALEY ALS Platform Trial and Radcliff studies who generously dedicated their time and effort. The HEALEY ALS Platform Trial was funded by the Sean M. Healey and AMG Center for ALS, the AMG Charitable Foundation, Tackle ALS, The ALS Association, ALS Finding a Cure®, Muscular Dystrophy Association, ALS One,

1 Austen data request form: <https://www.EverythingALS.org/available-data>

2 Repository: <https://github.com/EverythingALS/LEClinicallyMeaningfulDysarthriaALS>

3 Healey data request form: <https://www.massgeneral.org/assets/mgh/pdf/neurology/als/platform-trial-data-and-report-sharing-request-form.pdf>

4 PRO-ACT data request form: <https://ncr1.partners.org/ProACT/Account/Register>

Arthur M. Blank Family Foundation, I AM ALS, Tambourine and many other community fundraising initiatives and donors. Partial funding was also provided by UCB Pharma, Biohaven Pharmaceuticals, Clene Nanomedicine Inc., and Prilenia Therapeutics.

7 Corresponding Author

Correspondence to Julian Peller (julian@everythingals.org).

8 Ethics Declarations

8.1 Competing Interests

The authors declare the following competing interests. This work was conducted while ER and AT were paid consultants working at EverythingALS. ER is currently an employee of ITBA, and AT is currently an employee of SRI International. JP, MT, FS, GB, FA and DS are paid consultants to EverythingALS. JO and JSc are current employees of Bristol Myers Squibb. This work was conducted while RS was an employee of Bristol Myers Squibb. NN and OL are current employees and hold stock options in Regeneron Pharmaceuticals, Inc. OL serves as an unpaid scientific advisor to EverythingALS. HS has received personal compensation as an employee of Prilenia Therapeutics B.V. and holds stock options in the company. AK is an employee of J&J Innovative Medicine and holds stock options in the company. GS is an employee of Eli Lilly & Co. and holds company stock. AS has received grants and contracts for clinical research projects sponsored by the FDA, NIH/NIA, NIH/NINDS, The ALS Association, and ALS Finding a Cure Foundation, as well as study support from Shionogi, Inc., Biogen, and Amylyx. JSh is a consultant for Acurastem, Nura Bio, and has received research funds from the Healey Center. MCu is a consultant for Pasithea, Ono, Novartis, Roche, Inflectis, Coya, Otsuka, AC Immune, Aclipse, Immunity Pharm Ltd, Cytokinetics, Takeda, Biogen, Transposon, Quralis, Regeneron, AB Sciences, Locust Walk, Vector Y. and serves on the board of directors of Praxis Precision Medicine and has received institutional grants from Prilenia for Healey ALS Platform Trial. SP reports research grants from Amylyx Therapeutics, Revalesio, Eledon, Alector, UCB Pharma, Biohaven, Clene Nanomedicine, Prilenia Therapeutics, Seelos, Calico, Denali, Neurizon, NIH, DoD, the Muscular Dystrophy Association and consulting fees from Amylyx, Arrowhead, Biogen, BMS, Clene, Cytokinetics, Eikonizo, J&J, Merck, PharmAust, Prilenia, and

Sola. She has been a paid educational speaker for Medscape, PeerView, and i3Health. EF and LO serve as unpaid scientific advisors to EverythingALS. JB has received research support from Rapa Therapeutics, MT Pharma of America, ProJenX, Novartis, MDA, ALS Finding a Cure, the ALS Association, the Association for Frontotemporal Dementia and NINDS. JB has served as a consultant to Alexion Pharmaceuticals, Amylyx Pharmaceuticals, Biogen, MTPA, Regeneron Pharmaceuticals, Roon, Sanofi and Trace Neuroscience. JB also serves as an unpaid scientific advisor to EverythingALS. INB is the founder of EverythingALS.

8.2 Ethics Approval

All aspects of the design and conduct of the Everything ALS Speech Study (Austen) were approved by the Salus IRB. Every participant is presented with a digital written informed consent through a study portal and provides written documentation of informed consent prior to undergoing any study procedure, in accordance with the Declaration of Helsinki. All participants consented to inclusion of their study data (for example, coded identifiers, demographics, ALS history and outcome measures and speech samples) in a large speech database that can be accessed by researchers under appropriate data use agreements.

Radcliff study was approved by Salus IRB (Protocol number: 2022-09-RAD-PI42). It was conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

The HEALEY ALS Platform Trial was conducted in accordance with Good Clinical Practice guidelines of the International Conference on Harmonization and ethical principles of the Declaration of Helsinki. Protocol approval was provided for all trial sites by a central investigational review board as part of the Mass General Brigham Human Research Protection Program (HRPP). All participants were required to provide written informed consent before screening.

9 References

- Allison, K. M., Yunusova, Y., and Green, J. R. (2019). Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis. *American journal of speech-language pathology*, 28(1):96–107.
- Barnett, C., Green, J. R., Marzouqah, R., Stipancic, K. L., Berry, J. D., Korngut, L., Genge, A., Shoesmith, C., Briemberg, H., Abrahao, A., et al. (2020). Reliability and validity of speech & pause measures during passage reading in als. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 21(1-2):42–50.
- Berry, J. D., Paganoni, S., Carlson, K., Burke, K., Weber, H., Staples, P., Salinas, J., Chan, J., Green, J. R., Connaghan, K., et al. (2019). Design and results of a smartphone-based digital phenotyping study to quantify als progression. *Annals of clinical and translational neurology*, 6(5):873–881.
- Beukelman, D. R. and Mirenda, P. (2011). Augmentative and alternative communication: Supporting children and adults with complex communication needs. *Paul H. Brookes Publishing*.
- Bingham, I.N., Norel, R., Roitberg, E.G., Peller, J., Trevisan, M.A., Agurto, C., Merler, M., Shalom, D. E., Aguirre, F., Embon, I., et al. (2025). Listener effort measures clinically meaningful change of dysarthria in amyotrophic lateral sclerosis. *Brain Communications*, page fcaf232.
- Bleeker, S., Moll, H., Steyerberg, E. a., Donders, A., Derksen-Lubsen, G., Grobbee, D., and Moons, K. (2003). External validation is necessary in prediction research: A clinical example. *Journal of clinical epidemiology*, 56(9):826–832.
- Cabitza, F., Campagner, A., Soares, F., de Guadiana-Romualdo, L. G., Challa, F., Sulejmani, A., Seghezzi, M., and Carobene, A. (2021). The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, 208:106288.
- Chio, A., Mora, G., Lauria, G., Ghezzi, S., Mazzini, L., Cavallaro, T., Mandrioli, J., Cammarota, G., Restagno, G., Canosa, A., Soliveri, P., Barone, P., Calvo, A., D’Alfonso, S., Garofalo, E., Sasso, C.,

- Romoli, M., Lunetta, C., and Logroscino, G. (2016). Amyotrophic lateral sclerosis: Clinical management. *The Lancet Neurology*, 15(1):34–45.
- Collins, G. S., Dhiman, P., Ma, J., Schluskel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Martin, G. P., Moons, K. G., Van Smeden, M., et al. (2024a). Evaluation of clinical prediction models (part 1): from development to external validation. *bmj*, 384.
- Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., et al. (2024b). Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385.
- Dierckx, P. (1995). *Curve and surface fitting with splines*. Oxford University Press.
- Duffy, J. R. et al. (2012). *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., and Blau, H. M. (2019). Dermatologist level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.
- Hardiman, O., van den Berg, L. H., and Kiernan, M. C. (2017). Clinical diagnosis and management of amyotrophic lateral sclerosis. *Lancet Neurology*, 16(1):36–47.
- Harkey, B. A., Distefano, S., Pagliaro, J. A., Heyd, L., Chase, M., Igne, C., Yu, H., Sherman, A. V., Dagostino, D., Tustison, E., et al. (2025). Operational development and launch of an adaptive platform trial in amyotrophic lateral sclerosis: Processes and learnings from the first four regimens of the HEALEY ALS Platform Trial. *Muscle & Nerve*, 72(2):294-305.
- Harris, C., Mgbahurike, C., Kumar, N., and Yang, D. (2024). Modeling gender and dialect bias in automatic speech recognition. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15166–15184, Miami, Florida, USA. Association for Computational Linguistics.
- Johnson, S. A., Karas, M., Burke, K. M., Strackiewicz, M., Scheier, Z. A., Clark, A. P., Iwasaki, S., Lahav, A., Iyer, A. S., Onnela, J.-P., et al. (2023). Wearable device and smartphone data quantify als progression and may provide novel outcome measures. *NPJ Digital Medicine*, 6(1):34.

- Kiernan, M. C., Vucic, S., Cheah, B. C., Turner, M. R., Eisen, A., Hardiman, O., and Burrell, J. R. (2011). Amyotrophic lateral sclerosis. *Lancet*, 377(9769):942–955.
- Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.
- Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.
- McGarrigle, R., Johnson, D., Hughes, C., McDonagh, K., McDonald, D., and Kelly, S. (2017). Dysarthria in amyotrophic lateral sclerosis: A study of the role of listener effort in speech intelligibility. *Journal of Speech, Language, and Hearing Research*, 60:17–25.
- Merler, M., Agurto, C., Peller, J., Roitberg, E., Taitz, A., Trevisan, M. A., Navar, I., Berry, J. D., Fraenkel, E., Ostrow, L. W., et al. (2025). Clinical assessment and interpretation of dysarthria in als using attention based deep learning ai models. *npj Digital Medicine*, 8(1):1–10.
- Miotto, R., Wang, F., Wang, S., Jiang, X., and Sun, J. (2016). Deep learning for healthcare: Review, opportunities, and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246.
- Paganoni, S., Berry, J. D., Quintana, M., Macklin, E., Saville, B. R., Detry, M. A., Chase, M., Sherman, A.V., Yu, H., Drake, K. et al. (2022). Adaptive platform trials to transform amyotrophic lateral sclerosis therapy development. *Annals of neurology*, 91(2):165–175.
- Paganoni, S., Fournier, C. N., Macklin, E. A., Chibnik, L. B., Quintana, M., Saville, B. R., Detry, M. A., Vestrucci, M., Marion, J., McGlothlin, A., et al. (2025). Efficacy and Safety of Zilucoplan in Amyotrophic Lateral Sclerosis: A Randomized Clinical Trial. *JAMA network open*, 8(2):e2459058–e2459058.
- Paganoni, S., Harkey, B., Giacomelli, E., Cudkowicz, M., and Group, H. A. P. T. S. (2024). Lessons from the HEALEY adaptive platform trial in amyotrophic lateral sclerosis. *Nature Aging*, 4(11):1512–1515.

- Peller, J., Trevisan, M. A., Bujia, G., Aguirre, F., Shalom, D. E., Taitz, A., Henze, S., Bastola, S., Osik, J., Shewcraft, R. A., et al. (2025). Reliable monitoring of respiratory function with home spirometry in people living with amyotrophic lateral sclerosis. *Frontiers in Neurology*, 16:1588992.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.
- Riley, R. D., Archer, L., Snell, K. I., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., and Collins, G. S. (2024). Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *bmj*, 384.
- Stegmann, G. M., Hahn, S., Duncan, C. J., Rutkove, S. B., Liss, J., Shefner, J. M., and Berisha, V. (2021). Estimation of forced vital capacity using speech acoustics in patients with als. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 22(sup1):14–21.
- Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365.
- Van Calster, B., Steyerberg, E. W., Wynants, L., and Van Smeden, M. (2023). There is no such thing as a validated prediction model. *BMC medicine*, 21(1):70.
- Writing Committee for the HEALEY ALS Platform Trial, Berry, J. D., Maragakis, N. J., Macklin, E. A., Chibnik, L. B., Quintana, M., Saville, B. R., Detry, M. A., Vestrucci, M., Marion, J., McGlothlin, A., Stommel, E. W., Chase, M., Pothier, L., Harkey, B. A., Yu, H., Sherman, A., Shefner, J., Hall, M., Kittle, G., ... HEALEY ALS Platform Trial Study Group (2025). CNM-Au8 in Amyotrophic Lateral Sclerosis: The HEALEY ALS Platform Trial. *JAMA*, 333(13), 1138–1149.
- Writing Committee for the HEALEY ALS Platform Trial, Shefner, J. M., Oskarsson, B., Macklin, E. A., Chibnik, L. B., Quintana, M., Saville, B. R., Detry, M. A., Vestrucci, M., Marion, J., McGlothlin, A., Heiman-Patterson, T., Chase, M., Pothier, L., Harkey, B. A., Yu, H., Sherman, A. V., Hall, M.,

- Kittle, G., Berry, J. D., ... HEALEY ALS Platform Trial Study Group (2025). Pridopidine in Amyotrophic Lateral Sclerosis: The HEALEY ALS Platform Trial. *JAMA*, 333(13), 1128–1137.
- Writing Committee for the HEALEY ALS Platform Trial, Andrews, J., Paganoni, S., Macklin, E. A., Chibnik, L. B., Quintana, M., Saville, B. R., Detry, M. A., Vestrucci, M., Marion, J., McGlothlin, A., Young, E., Chase, M., Pothier, L., Harkey, B., Yu, H., Sherman, A., Shefner, J., Hall, M., Kittle, G., ... Babu, S. (2025). Verdiperstat in Amyotrophic Lateral Sclerosis: Results From the Randomized HEALEY ALS Platform Trial. *JAMA neurology*, 82(4), 333–343.
- Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Albu, E., Arshi, B., Bellou, V., Bonten, M. M., et al. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369.
- Yorkston, K. M., Beukelman, D. R., and Strand, E. A. (1996). Management of motor speech disorders in children and adults. *Thomson*.
- Yunusova, Y., Green, J. R., Wang, J., Pattee, G., and Zinman, L. (2011). A protocol for comprehensive assessment of bulbar dysfunction in amyotrophic lateral sclerosis (als). *Journal of visualized experiments: JoVE*, (48):2422.