Digital Biomarkers

Research Article

Digit Biomark 2022;6:71–82
DOI: 10.1159/000525698

Received: March 23, 2022
Accepted: June 19, 2022
Published online: July 21, 2022

Reliability of Automatic Computer Vision-Based Assessment of Orofacial Kinematics for Telehealth Applications

Leif Simmatis[a, b] Carolina Barnett[c, d, e] Reeman Marzouqah[a] Babak Taati[b, f, g] Mark Boulos[d, h] Yana Yunusova[a, b, h]

[a]Department of Speech-Language Pathology, University of Toronto, Toronto, ON, Canada; [b]KITE-Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada; [c]Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, ON, Canada; [d]Division of Neurology, Department of Medicine, University of Toronto, Toronto, ON, Canada; [e]University Health Network, Toronto, ON, Canada; [f]Department of Computer Science, University of Toronto, Toronto, ON, Canada; [g]Institute of Biomedical Engineering, University of Toronto, Toronto, ON, Canada; [h]Sunnybrook Health Sciences Centre, Toronto, ON, Canada

Keywords
Digital biomarkers · Digital devices · Mobile technology

Abstract

Introduction: Telehealth/remote assessment using readily available 2D mobile cameras and deep learning-based analyses is rapidly becoming a viable option for detecting orofacial and speech impairments associated with neurological and neurodegenerative disease during telehealth practice. However, the psychometric properties (e.g., internal consistency and reliability) of kinematics obtained from these systems have not been established, which is a crucial next step before their clinical usability is established. ***Methods:*** Participants were assessed in lab using a 3 dimensional (3D)-capable camera and at home using a readily-available 2D camera in a tablet. Orofacial kinematics was estimated from videos using a deep facial landmark tracking model. Kinematic features quantified the clinically relevant constructs of velocity, range of motion, and lateralization. In lab, all participants performed the same oromotor task. At home, participants were split into two groups that each performed a variant of the in-lab task. We quantified within-assessment consistency (Cronbach's α), reliability (intraclass correlation coefficient [ICC]), and fitted linear mixed-effects models to at-home data to capture individual-/task-dependent longitudinal trajectories. ***Results:*** Both in lab and at home, Cronbach's α was typically high (>0.80) and ICCs were often good (>0.70). The linear mixed-effect models that best fit the longitudinal data were those that accounted for individual- or task-dependent effects. ***Discussion:*** Remotely gathered orofacial kinematics were as internally consistent and reliable as those gathered in a controlled laboratory setting using a high-performance 3D-capable camera and could additionally capture individual- or task-dependent changes over time. These results highlight the potential of remote assessment tools as digital biomarkers of disease status and progression and demonstrate their suitability for novel telehealth applications.

Karger@karger.com
www.karger.com/dib

Correspondence to:
Leif Simmatis, leif.simmatis@uhn.ca
Yana Yunusova, yana.yunusova@utoronto.ca

Introduction

Individuals with neurological disorders frequently experience impairments in orofacial function, as well as speech and swallowing disorders. Standard clinical practice for decades has focused on clinician-administered assessments, either conducted informally or using tools such as the House-Brackmann scale [1] for the assessment of facial nerve function and the Frenchay Dysarthria Assessment (FDA) for the assessment of motor speech disorders and associated orofacial impairments [2]. However, these tools tend to be perception-based and can therefore introduce human error. For example, the House-Brackmann scale suffers from substantial differences in agreement depending on the severity of facial paresis in a given patient [3].

Instrumental assessments of orofacial function such as electromagnetic articulography and multicamera optical tracking can precisely and reliably measure orofacial movement across various neurological diseases. Kinematics obtained using these instrumental tools have been used to identify markers of early disease and disease progression in patients with amyotrophic lateral sclerosis [4–6], estimate medication state in Parkinson's disease [7], and quantify stroke-related impairments in facial motor control [8]. Although highly accurate [9–12], these orofacial assessment systems require specialized (and therefore expensive) hardware and software, as well as substantial setup, post-processing, and measurement time. Furthermore, they are beyond the reach of home-based use, which further limits their utility for regular at-home monitoring and/or assessing individuals with ambulation challenges.

Recently, inexpensive technologies such as high-resolution 3-dimensional (3D) cameras paired with sophisticated artificial intelligence methods (e.g., deep neural networks) have been introduced as means of automating orofacial assessment in a laboratory setting [13]. Lab-quality cameras with 3D capabilities such as the Intel® RealSense™ can be used to accurately capture impairments associated with disease [14]. Although these cameras are cumbersome to use and thus are not feasible for home-based use, high-quality 2D cameras in consumer electronics (e.g., smartphones and tablets) show promise for enabling high-quality orofacial assessment remotely from patients' homes. Recent research has demonstrated that that artificial intelligence-enabled remote assessment using 2D cameras can detect neurological impairment [15] and frequent at-home assessments of individuals with neurodegenerative disease can enhance disease monitoring [16]. There is an urgent need to explore the measurement properties of 2D camera systems in the context of remote assessment.

The creation and deployment of digital health tools has been recently delineated using frameworks such as V3 [17] and the FDA BEST model [18]. The V3 framework consists of verification, analytical validation, and clinical validation stages. Verification ensures that sensors produce values within a target tolerance, analytical validation ensures that novel methods are comparable to validated gold standards, and clinical validation ensures that putative digital biomarkers reflect clinical constructs of interest. Previous studies included an analytical validation of 3D cameras in comparison to relevant ground truths [19], and early clinical validation of 3D cameras has been performed in the context of detecting neurological impairments [20]. Our study aimed to contribute to the analytical validation of a 2D camera as an essential step in its eventual clinical adaptation. For home-based recording technologies, the process of analytical validation includes comparing their psychometric properties to those of a lab-based 3D-capable camera.

Desirable measurement properties for biomarkers include good internal consistency, good reliability, and the ability to track individual- or task-dependent performance over time [21]. Remote assessment is particularly amenable to these types of analyses, given both the ease of collecting data at home compared to in lab [16], and the increased statistical power associated with performing multiple repeated tests [22]. Thus, we aimed to establish psychometric characteristics in markerless facial landmark tracking performed remotely using a 2D camera. Specifically, we addressed the research questions of (1) whether orofacial kinematics estimated from home-based assessments had comparable internal consistency to those from in-lab data, (2) whether home-based orofacial kinematics had comparable test-retest reliability to those gathered in-lab, and (3) whether longitudinal home-based facial kinematics were sensitive to individual trends and group differences. We hypothesized that data collected at home would be comparable in reliability within and across multiple assessments relative to the in-lab assessments. We further hypothesized that our approach would be able to detect differences in individual longitudinal trajectories of orofacial kinematics across different tasks. The primary contribution of this article is a step towards analytical validation of home-based video assessment of orofacial function compared to lab-based gold standards, which will enable further development of remote orofacial and speech assessment technologies.

Methods

Participants and Clinical Assessment

We assessed 13 individuals with chronic stroke recruited from the stroke clinic at Sunnybrook Health Sciences Centre (Toronto, ON, Canada) as part of a study of the effects of exercise-based intervention on poststroke obstructive sleep apnea. Inclusion criteria were presence of grossly typical craniofacial structures, as well as absence of oral apraxia, apraxia of speech, aphasia, significant depressive symptoms, and other neurological/neurodegenerative conditions. These individuals had normal levels of speech function and were assumed to be unlikely to change over time. Furthermore, given that these individuals had scores of 0 on the FDA, indicating normal oromotor function, we did not expect clinically relevant changes in oromotor function over the duration of this study.

At baseline, participants were assessed using the Montreal Cognitive Assessment (MoCA) [23] to document cognitive status, FDA to assess oromotor and speech functions [2], and in-lab video-based orofacial kinematic assessment to instrumentally document range/velocity of movement and lateralization during oromotor tasks. The in-lab orofacial assessment was conducted using a 3D-capable camera (Intel RealSense SR300), although we only analyzed videos subsequently in 2D. Participants returned to the lab for the same assessment at post-training (6 weeks after baseline) and retention (4 weeks after post-training).

After the baseline in-lab assessment, participants were randomly allocated (1:1) to perform two different sets of tasks at home: (1) Group 1 performed simple range of motion (ROM) tasks ("sham tasks") and (2) Group 2 performed exercises that focused on the oropharyngeal muscles ("OPE tasks"). Both groups completed the exercises at home using a custom application called OPEX (OroPharyngeal EXercise) deployed on a Samsung Galaxy Tab A7 tablet. The tablet was stabilized on a stand and participants' faces were video recorded during the exercises using the built-in front camera. Participants were trained on the app to ensure good quality of video recordings. Follow-up phone calls and in-person/Skype visits were scheduled to download data, provide retraining if needed, and troubleshoot technical issues. The at-home data collection routine consisted of up to 60 assessments (maximum of 2 per day for 30 days spread over 6 5-day weeks).

This study was approved by the Research Ethics Boards at Sunnybrook Research Institute. Written and informed consent was obtained by all the participants in accordance with the Declaration of Helsinki.

Orofacial Assessment Tasks

Participants in this study performed several tasks in lab and at home; however, for the purposes of the present analysis, we focused only on the tasks that required vertical movement of the lower jaw and lip. In the in-lab condition, all participants performed a jaw opening and closing task in which they were instructed to open and close their mouth as fast and as far as possible, for approximately 5 movement repetitions per assessment (range 3–5 were observed in the data). In the at-home condition, participants were split into two groups to perform two versions of this task. Individuals performing the sham task were instructed to open the mouth as wide as comfortably possible, moving the jaw at a comfortable rate, and then close it again, returning to a neutral expression. The OPE task was a variation of the sham task, in which participants opened their mouth while pressing their tongue with maximum effort onto the roof of their mouth, effectively anchoring their tongue during mouth opening. To complete this task, while the mouth was still closed, participants pressed their tongue onto the roof of their mouth with maximal effort. They then maintained this upward pressure using their tongue while opening their jaw as far as was comfortably possible and then closing it again. The OPE task was designed to increase tongue strength to assist with obstructive sleep apnea symptoms, but we would not expect it to dramatically influence orofacial movement velocities after sufficient training. At home, the typical number of repetitions included was higher than in-lab; participants were instructed to perform approximately 10 repetitions, and 6–8 repetitions were typically observed in the data used for the analysis. The reasons for the typical number of observed repetitions differing from the instructed number ($n < 10$) include some repetitions being poor quality (e.g., partially obscured face, obstruction, or excessive movement), or in some cases participants did not complete 10 repetitions at all.

Orofacial Kinematic Estimation

Orofacial kinematics was estimated from video using a deep learning model along with automated signal processing routines. See the online supplementary Information (for all online suppl. material, see www.karger.com/doi/10.1159/000525698) for an in-depth technical description of our kinematic feature extraction process. Briefly, full-length videos were preprocessed manually to remove sections that deviated from task protocol (e.g., having off-camera conversations). Videos were then analyzed using a deep learning facial alignment model called the facial alignment network [24], which predicted 68 facial landmarks per frame of video; see Figure 1 for locations of all landmarks and relevant facial distances. Landmarks of interest were: (1) one located at the vermilion border of the lower lip at midline (LL), (2) landmarks at the left and right corners of the mouth (ML and MR, respectively), and (3) a reference landmark that was used to account for head position changes throughout the video (REF) [14]. The positions of these landmarks over time were smoothed using a 4th order low-pass Butterworth filter. Distances were normalized by the intercanthal distance, and then converted to units of millimeters instead of pixels using an estimated true intercanthal distance of 33 mm [25]. Each processed timeseries of landmark positions consisted of multiple movement repetitions, so individual repetitions were captured using a peak-detection technique [5].

Extracted kinematics included velocity of the LL landmark (LL-velocity), ROM, and lateralization. LL-velocity was calculated as the average vertical velocity over the window of time between movement onset and offset (values reported in mm/s). Lateralization was calculated by measuring the horizontal excursion of the LL point from the midpoint of the line between ML and MR (values reported in mm). Finally, ROM was derived by calculating the difference between the maximum and minimum vertical positions of LL per movement repetition (values reported in mm). LL-velocity and lateralization (not ROM) were quantified separately for opening and closing mouth movements.

Statistical Analysis

To address the question of whether kinematics were consistent across multiple repetitions within assessment, Cronbach's α was used [26]. Here, α was calculated across movement repetitions for all participants. In-lab, α was calculated for each of the 3 assess-

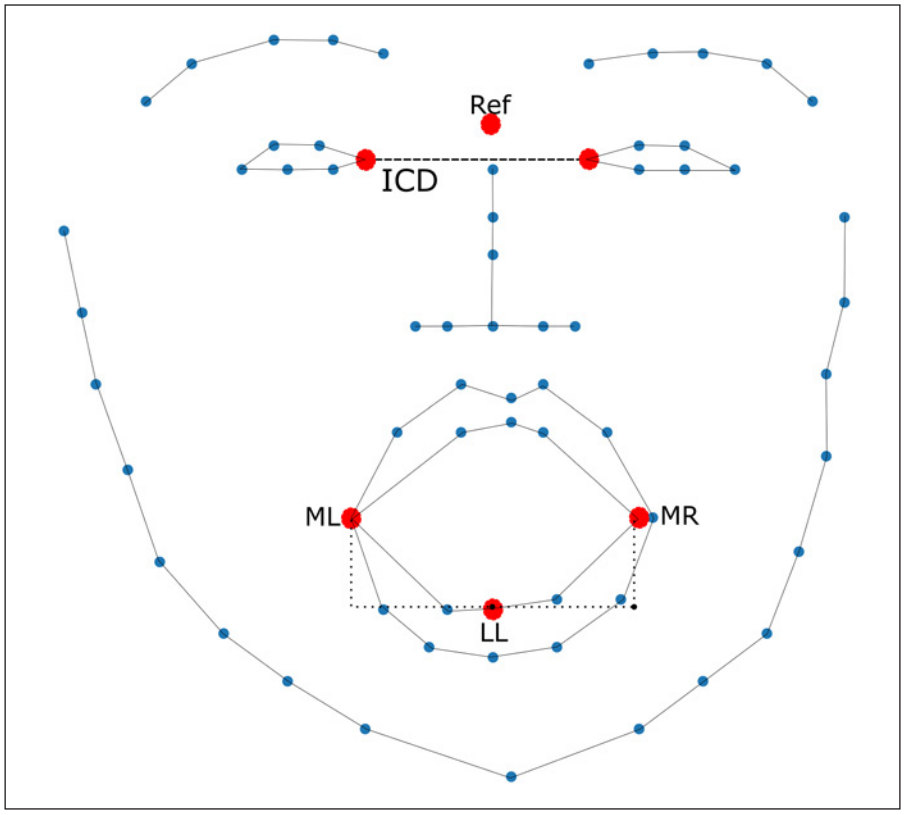


Fig. 1. An image of predicted facial landmarks for a single frame (n = 68). The key landmarks used for calculations are large circles. The line labeled ICD was used as a distance normalization line. Point LL was tracked for speed calculations. Lateralization was calculated as the signed horizontal deviation of point LL from the midpoint of the two points ML and MR. ICD, intercanthal distance.

Table 1. Summary statistics for OPE and sham groups

N	10 (lab)	13 (home)
Sex (*n*, F)	4	4
Age, median [IQR]	65.5 [11.0]	66.0 [11.0]
MoCA		
Median [IQR]	26.5 [5.8]	27.0 [6.0]
N < 23	4	4
Previous TIA	4	4
FDA score >0 at enrollment	0/10	0/13
Time since stroke, weeks		
Median [IQR]	133.9 [103.6]	178.4 [183.9]
Range	57.3–267.9	15.4–379.6

M, male; F, female; MoCA, Montreal Cognitive Assessment; IQR, interquartile range; OPE, oropharyngeal exercise group; TIA, transient ischemic attack; FDA, Frenchay Dysarthria Assessment.

ments. At home, participants completed a variable number of assessments, but most completed ≤20 in total; therefore, at-home α was calculated for the 1st assessment, the 3rd assessment (as an indicator of internal consistency early in follow-up), and the 20th assessment (the last assessment before substantial attrition was observed). Both in lab and at home, we quantified α at multiple timepoints to account for potential changes in internal consistency over time. Within-trial standard deviations (SDs), pooled within-trial SDs, and across-trial SDs were also calculated for each of the kinematic measures for the sessions described above (see online suppl. Table 1).

Assuming no change in function over time, a reliable measurement tool should be able to consistently estimate metrics and distinguish between different individuals [27]. Therefore, we quantified the test-retest reliability of orofacial kinematics across multiple assessments using the intraclass correlation type (2,1) (ICC), which assumes a random selection of raters [28]. In-lab ICCs were calculated for two cases: (A) all three in-lab assessments and (B) the last two in-lab assessments. Analyses were stratified in this way to account for any potential learning effects form the at-home exercise paradigm, which was completed between the 1st and 2nd in-lab assessments. For at-home assessments, two cases were also considered: (A) the first three at-home assessments, and (B) the last three at-home assessments, in order to determine if reliability changed as time went on. Three assessments were used each time for comparability to in-lab ICCs, avoiding spurious changes in ICC due to sample size [29].

We employed linear mixed-effects regression (LME) models [30] to capture individual- and task-dependent differences in orofacial kinematics over time and augment our descriptions of the reliability of orofacial kinematics. LME models were chosen because they are robust to missing/unevenly spaced data, abnormally distributed data and are able to accommodate a variety of model specifications. We only analyzed at-home data using LME models because the at-home data had enough repetitions to reasonably capture individual trends. LME models were fit to the data from

Table 2. Summary of within- and between-test reliability

Assessment	ROM	Velocity		Lateralization	
	(stat, CI 95%)	open (stat, CI 95%)	close (stat, CI 95%)	open (stat, CI 95%)	close (stat, CI 95%)
Lab					
Alpha					
1st	0.96 (0.89, 0.99)	0.88 (0.66, 0.97)	0.94 (0.83, 0.98)	0.88 (0.65, 0.97)	0.96 (0.89, 0.99)
2nd	0.97 (0.91, 0.99)	0.85 (0.55, 0.96)	0.84 (0.52, 0.96)	0.96 (0.90, 0.99)	0.98 (0.94, 0.99)
3rd	0.55 (–0.40, 0.89)	0.77 (0.27, 0.94)	0.85 (0.52, 0.96)	0.95 (0.85, 0.99)	0.97 (0.91, 0.99)
ICC					
Last 2	0.49 (–0.16, 0.84)	0.67* (0.12, 0.91)	0.05 (–0.57, 0.63)	0.70* (0.16, 0.91)	0.61* (0.01, 0.88)
All 3	**0.59 (0.21, 0.86)**	**0.70 (0.36, 0.90)**	0.20 (–0.15, 0.65)	**0.64 (0.28, 0.88)**	**0.61 (0.24, 0.87)**
Home					
Alpha					
1st	0.94 (0.86, 0.98)	0.96 (0.92, 0.99)	0.95 (0.89, 0.98)	0.96 (0.91, 0.99)	0.90 (0.78, 0.97)
3rd	0.84 (0.66, 0.95)	0.92 (0.83, 0.97)	0.89 (0.77, 0.96)	0.88 (0.74, 0.96)	0.84 (0.65, 0.94)
20th	0.92 (0.82, 0.98)	0.95 (0.89, 0.99)	0.96 (0.90, 0.99)	0.93 (0.85, 0.98)	0.90 (0.76, 0.97)
ICC					
First 3	**0.74 (0.47, 0.91)**	**0.74 (0.47, 0.91)**	**0.64 (0.31, 0.86)**	0.31* (–0.04, 0.68)	**0.42 (0.07, 0.75)**
Last 3	**0.79 (0.52, 0.94)**	**0.87 (0.67, 0.96)**	**0.97 (0.91, 0.99)**	**0.73 (0.41, 0.92)**	**0.66 (0.31, 0.89)**

Stat, relevant statistic (either ICC or alpha, depending on table row); CI 95%, 95% confidence interval; ICC significance, **bold** values are associated with p values <0.0025; values with an asterisk*, $0.0025 < p < 0.05$.

first 20 assessments, and four different LME model structures were compared, given that most people in the study completed at least 20 sessions before substantial attrition was observed. The following models were assessed: "Model 1" (random intercept per-individual and no slope with respect to time); "Model 2" (random intercepts per-individual with a fixed slope with respect to time); "Model 3" (random intercepts per-individual as well as random slope with respect to time per-individual); and "Model 4" (random intercepts per-individual, random influence of time per-individual, and random influence of time per-task). Models 1–3 were compared pairwise using analysis of variance (ANOVA). Models 3 and 4 were compared to investigate specifically the value of stratifying by group/task over and above stratifying by individual. To avoid many model comparisons, no other contrasts with Model 4 were considered (e.g., Model 1 and Model 4 were not compared). It would also be of little practical value to compare some of the models (e.g., Model 1 and Model 4). Marginal- and conditional R^2 for LME models were reported; model fits were compared using Akaike Information Criterion (AIC) and analysis of variance (ANOVA). The absolute magnitudes of the individual regression slopes were contrasted to investigate task effects in kinematic tracking. LME models were implemented using the R (version 4.0.3) lme4 package [31], run in a Python environment using the rpy2 package.

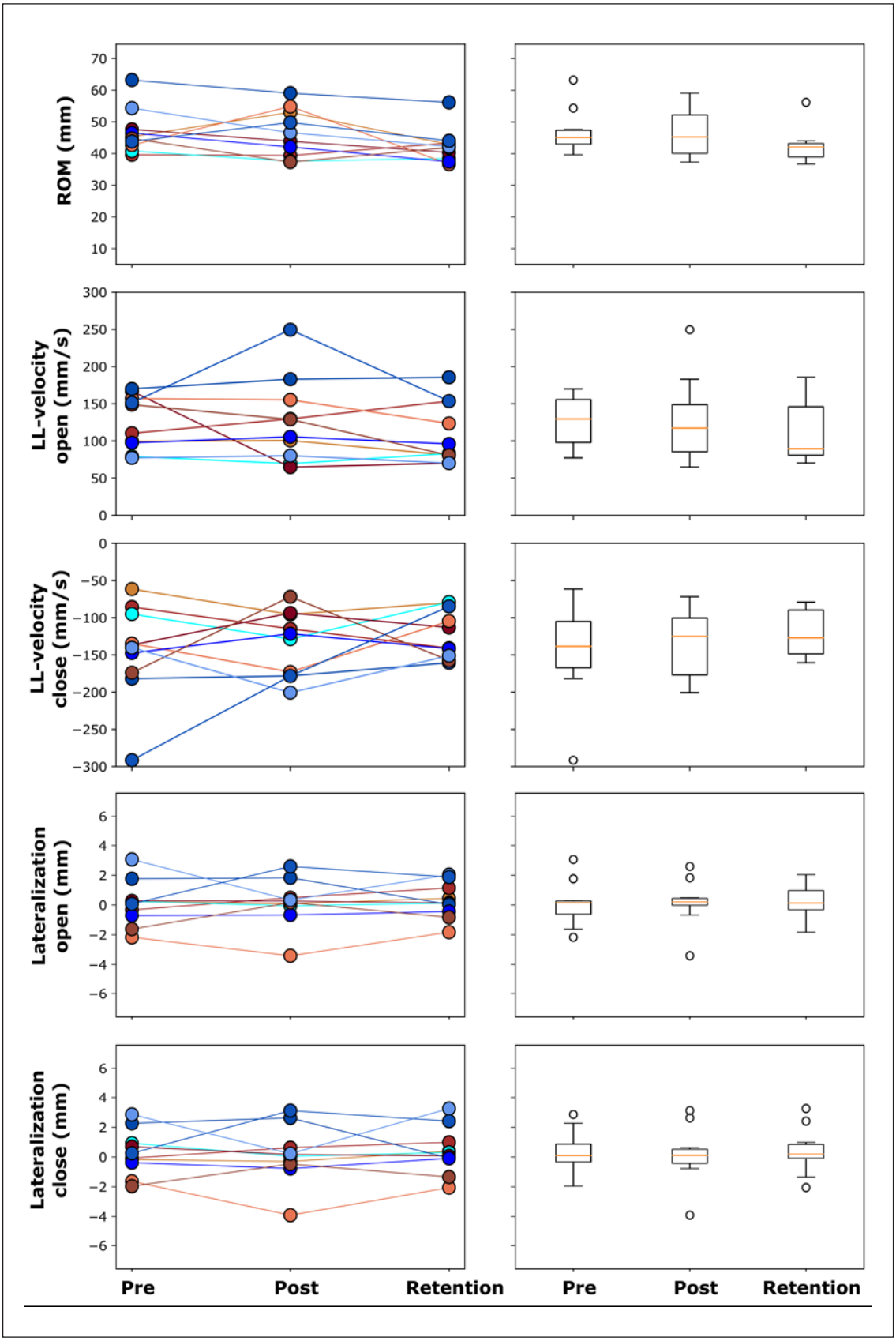
Results

Participants, Demographics, and Overview of Kinematics

The data from 10 individuals with complete datasets across 3 assessments were included in the analyses of in-lab data, and 13 were included for at-home data. The 10 individuals in the in-lab case are a subset of the 13 who completed the at-home assessment. Table 1 provides a summary of demographic and clinical data. For the in-lab set, 6 individuals were male and 4 were female (at-home: 9 and 4, respectively), the median (IQR) age was 66.0 (11.0) years (at-home: 66.0 [11.0]), and the median (IQR) time since stroke was 89.3 (114.8) weeks (at-home: 178.4 [183.9] weeks). Most (6/10 in-lab; 9/13 at-home) individuals had MoCA > 23 (the average measured previously from a large, ecologically valid cohort [32]). Remaining individuals had scores 14–20 but were functionally able to perform the tasks in lab and at home independently, as judged by recording quality and task performance. At enrollment, all individuals scored a 0 on the FDA, indicating normal orofacial and oropharyngeal muscle function and no dysarthria.

Internal Consistency for In-Lab and At-Home Assessments

Consistency across multiple trials within each assessment was generally high across all measures both in lab and at home. See Table 2 for a summary of Cronbach's α values. For in-lab data, α ranged between 0.55 and 0.98 across all sessions, and 13/15 measures had $\alpha \geq 0.80$, indicating high internal consistency. The exceptions were the 3rd assessment of lower lip (LL)-velocity (open) (α =



2

(For legend see next page.)

0.77), and the 3rd assessment of ROM (α = 0.55). At home, α values ranged between 0.84 and 0.96, i.e., all 15 measures had α ≥ 0.80, suggesting very high internal consistency.

Test-Retest Reliability of In-Lab and At-Home Assessments

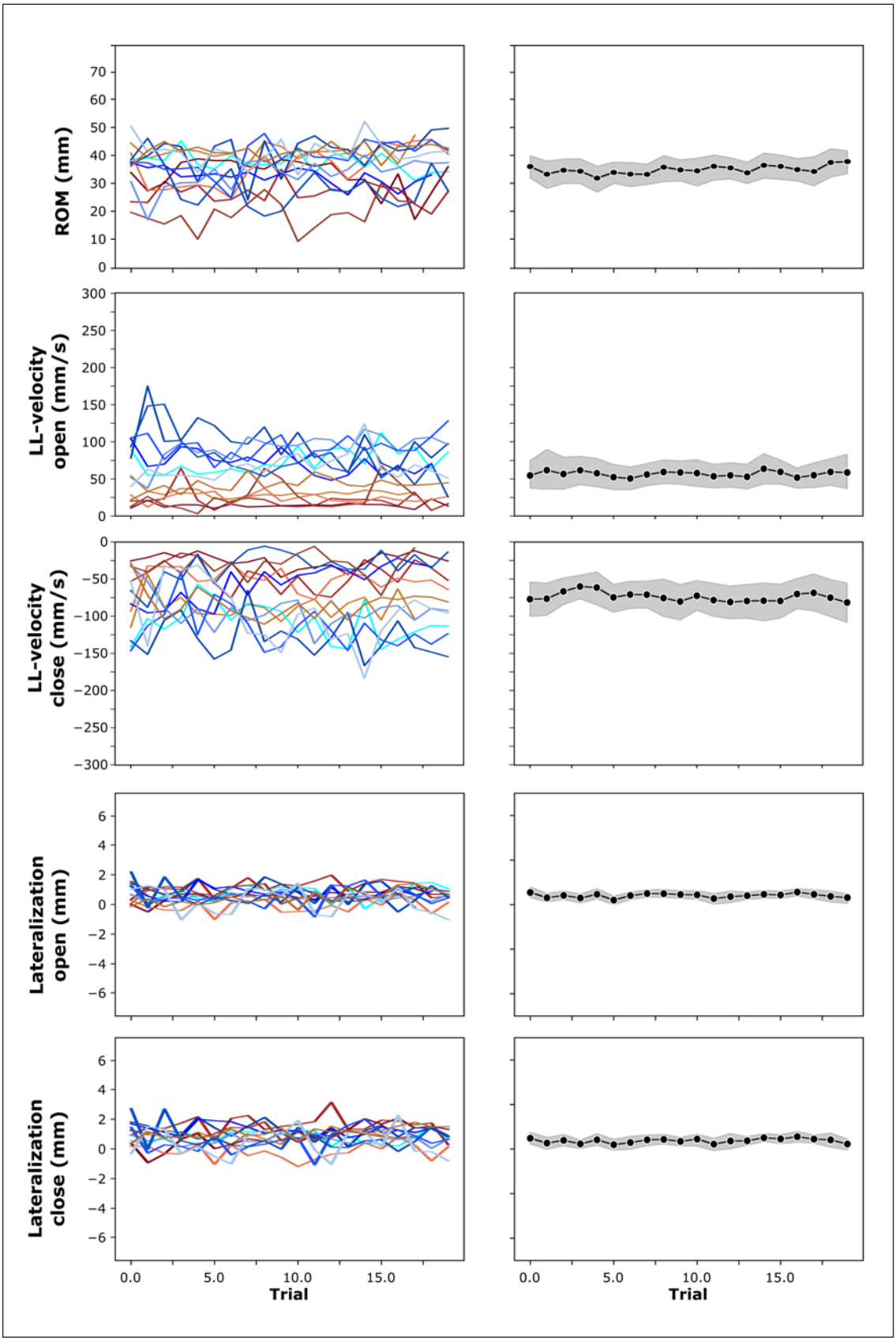
As shown in Table 2, reliability between assessments was generally moderate-to-good, for both in-lab and at-home data. In-lab, ICCs ranged between 0.05 and 0.72. Out of 10 in-lab ICCs, 4 were statistically significant after correction for multiple comparisons (i.e., $p < 0.0025$), and another 3 approached statistical significance. The lowest in-lab ICCs (values of 0.05 and 0.21, respectively) were found in the LL-velocity (close); the highest ICC was found in lateralization (open) across the last 2 assessments (ICC = 0.72). For at-home data, ICCs ranged between 0.30 and 0.97. Eight out of ten at-home ICCs were significant after correction for multiple comparisons, and the remaining two approached significance. The lowest ICC was found in lateralization (open), across the first 3 assessments (ICC = 0.30), whereas the highest ICC was found in LL-velocity (close) across the last 3 assessments (ICC = 0.97). Figures 2 and 3 provide graphical depictions of change over time, both for in-lab data and at-home data, respectively, with sham and OPE tasks color-coded in shades of blue and red, respectively. For in-lab data, consistency across assessments was visibly highest for

Table 3. Comparison of LME model fits for at-home data

Kinematic measure	Models	χ^2 (d.f.)	*p* value	R^2 (mar., con.)	Mean |β| (sham, OPE)
ROM	Model 1, Model 2	0.093 (1)	0.760	–	–
	Model 1, Model 3	21.999 (3)	**6.526e−5**	–	–
	Model 2, ***Model 3***	21.906 (2)	**1.751e−5**	(0.012, 0.759)	(0.376, 0.219)
	Model 3, Model 4	0.000 (3)	1.000	–	–
LL-velocity (open)	Model 1, Model 2	1.896 (1)	0.169	–	–
	Model 1, Model 3	38.018 (3)	**2.802e−8**	–	–
	Model 2, Model 3	36.122 (2)	**1.433e−8**	–	–
	Model 3, ***Model 4***	14.896 (3)	0.002*	(<1.0e−4, 0.903)	(1.982, 0.471)*
LL-velocity (close)	Model 1, Model 2	0.004 (1)	0.947	–	–
	Model 1, Model 3	21.108 (3)	**9.995e−5**	–	–
	Model 2, ***Model 3***	21.104 (2)	**2.614e−5**	(0.006, 0.763)	**(1.783, 0.343)**
	Model 3, Model 4	1.635 (3)	0.652	–	–
Lateralization (open)	***Model 1***, Model 2	0.274 (1)	0.601	(<1.0e−4, 0.269)	n/a
	Model 1, Model 3	1.177 (3)	0.759	–	–
	Model 2, Model 3	0.903 (2)	0.637	–	–
	Model 3, Model 4	0.000 (3)	1.000	–	–
Lateralization (close)	Model 1, Model 2	2.253 (1)	0.133	–	–
	Model 1, Model 3	10.700 (3)	0.013*	–	–
	Model 2, ***Model 3***	8.447 (2)	0.015*	(0.006, 0.428)	(0.011, 0.015)
	Model 3, Model 4	0.000 (3)	1.000	–	–

Model 1, intercept only; Model 2, random intercept, fixed slope; Model 3, random intercept, random slope; * asterisks, $p < 0.05$ but not after multiple comparisons correction; **bolded** values, significance after multiple comparisons correction; ***Bolded and italicized*** models, those that were selected as optimal after comparison; D.f., degrees of freedom used in χ^2 test; ROM, range of motion; R^2 marginal (mar.) and conditional (con.) values are reported only for the best models; remaining rows are filled with dashes (–). |β|, absolute magnitude of individualized regression slopes. n/a, that there were no individualized slopes considered.

Fig. 2. Line plots of individual assessments for LL-velocity, lateralization, and ROM metrics during in-lab assessment (left side of each panel), and boxplots showing the same data in summarized form (right side of each panel). Individuals that performed the sham task are in shades of blue and individuals that performed the OPE task are in shades of red.



3

(For legend see next page.)

ROM and LL-velocity (open), supporting the numeric results for ICC.

At-Home Individual- and Task-Dependent Longitudinal Tracking of Kinematics

Table 3 presents summaries of model comparisons and model fits for at-home data, as well as summaries of slope magnitudes for cases where Model 3 and Model 4 were chosen. The majority (3/5) of the optimal LMEs were Model 3 (i.e., random effect of time and a random intercept, without random effect of group/task membership). The kinematic variables for which Model 3 was the best fit were LL-velocity (close), lateralization (close), and ROM. Lateralization (open) was the only case in which no subsequent changes beyond Model 1 (i.e., fixed slope and fixed intercept) provided a significantly better fit to the data. Model 4 was the best-fitting model only for LL-velocity (open) in the at-home condition; the sham task (blue) and OPE task (red) have subjectively differing means and trajectories. Furthermore, in the LL-velocity (open) and LL-velocity (close) models, the mean absolute individualized regression slopes between sham and OPE tasks were significantly different. However, only the LL-velocity (close) case was significant after multiple comparisons correction.

Discussion

In line with V3 (Verification, analytical Validation, and clinical Validation), COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments), and BEST (Biomarkers, EndpointS, and other Tools) frameworks [18, 33, 34], biomarkers and testing instruments should be reliable, internally consistent, and sensitive to group differences. We demonstrated that orofacial kinematics gathered remotely using consumer-grade cameras possessed these favorable properties, and furthermore that they could identify individual- and task-dependent differences over time. Specifically, we showed that (1) within-test consistency and test-retest reliability of remotely gathered kinematics were at least as good as they were for in-lab kinematics, and (2) remotely gathered orofacial kinematics could capture individual- and task-dependent differences in performance over time. In doing so, we demonstrated that information gathered at home from 2D cameras had comparable value to that from in-lab 3D cameras.

Fig. 3. Line plots of individual assessments for LL-velocity, lateralization, and ROM metrics for at-home assessments. Plots in the right-hand column are summarized versions of the data in the left-hand column (mean ± 1 standard deviation). Individuals that performed the sham task are in shades of blue and individuals that performed the OPE task are in shades of red.

Intraindividual variability and test-retest reliability are both important for evaluating motor performance, as well as capturing disease state [35–40], measuring treatment effects [41], and ultimately for identifying meaningful behavioral biomarkers [42]. The present study demonstrated that video-based orofacial kinematics possess these properties, and that they are at least as good for remotely collected data as for those gathered in-lab. Furthermore, we demonstrated that estimates of within- and between-test reliability were stable over time. These findings demonstrate the viability of using remote assessment methods in place of lab-based assessment methods, even for precision kinematic assessment. The lowest reliability estimate that we observed was in the in-lab setting, for LL-velocity (close); this was potentially caused by small sample size, low repetition number, and lack of heterogeneity across individuals. In contrast to the at-home condition, participants in lab performed all the same task, which would have attenuated interindividual variability (a factor in reliability). It is possible, given these varying results, that LL-velocity (close) may not be an ideal measure to use in the future. It is likely that focusing on mouth opening movement specifically, or the whole open/close movement sequence, would be more informative. Collectively, our results suggested that longitudinal at-home (remote) orofacial assessment could be valuable for tracking disease progression, as well as recovery (or maintenance of function) following pharmaceutical and/or behavioral interventions [43, 44].

Remote longitudinal assessment tools for orofacial and speech applications have several clinical and research benefits. From a research perspective, these home-based assessments can be used to detect impairments associated with neurodegenerative disease that, with further development, could aid in clinical decision-making [15]. Home-based assessment also enables in-depth quantification of stability or change over time in individuals with neurological diseases due to disease progression or intervention [5, 7, 45]. Longitudinal assessments typically have greater statistical power than cross-sectional data [46, 47], and longitudinal lab-based orofacial assessments have been shown to distinguish between disease subgroups [35]. Leveraging the potential of remote assessment tools to easily collect longitudinal data will be important for the development of future orofacial assessment technologies.

Furthermore, from a clinical perspective, telehealth/remote methods with frequent follow-ups can enable improved access to care, including clinical trials of pharmaceutical and behavioral interventions, close monitoring of conditions in natural settings and with minimal time constraints (e.g., time of the day). By validating the reliability of orofacial kinematics gathered in lab compared to at home, the present study represents an important step forward in the technology development process. Further work will be required to expand these validation efforts, particularly into the domain of clinical validity, continuing to use the frameworks such as V3 [17].

Limitations

Our study has some limitations that should be addressed in future research. First, our sample size was relatively small, which may impact the generalizability of our results. Nevertheless, most of our estimates of internal consistency and across-assessment reliability had small confidence intervals, suggesting that for these purposes our sample size was reasonable. Second, some of our participants had low scores on the MoCA; however, inspection of the videos for these participants indicated that they performed our relatively simple oromotor tasks correctly, and so we included them because the current tasks did not test cognition. We were unable to fully compare orofacial kinematics across in-lab and at-home conditions because the task constraints were different between settings. For future studies that validate the use of remote collection techniques – both web-based and mobile data collection apps – oromotor tasks should be more varied and fully consistent.

Conclusions

We demonstrated that objective orofacial kinematics obtained from remotely collected videos were reliable within assessment, stable across repeated assessments, and could capture interindividual heterogeneity. Additionally, we demonstrated that remote assessments and 2D consumer-grade cameras provided data that were comparably reliable to that obtained in a controlled laboratory setting using 3D camera equipment. The present results suggest that these orofacial kinematic metrics could be valuable instrumental, objective, and clinically relevant digital biomarkers of motor speech function and could be used in the future to augment clinical decision-making via at-home monitoring/virtual/telehealth for individuals with neurological disorders.

Acknowledgments

We would like to sincerely thank Dr. Madhura Kulkarni for her technical assistance, help data curation, and help with organization.

Statement of Ethics

This study protocol was reviewed and approved by the Research Ethics Board at Sunnybrook Research Institute (Toronto, ON, Canada), approval number 1726. Written and informed consent was obtained by all the participants in accordance with the Declaration of Helsinki.

Conflict of Interest Statement

The authors declare no competing interests.

Funding Sources

This study was supported by a Canadian Partnership for Stroke Recovery (CPSR) Collaborative Grant, a Michael J. Fox Foundation and Weston Brain Institute Computational Science Fellowship, a National Institutes of Health R01 grant (NIH R01DC017291), and an Age-Well NCE Trainee Grant. The fundings agencies had no role in study design, nor the collection, analysis, or interpretation of study data.

Author Contributions

Leif Simmatis was responsible for conceptualization, data analysis, drafting of the original manuscript, and editing. Carolina Barnett edited the draft. Reeman Marzouquah edited the draft. Babak Taati edited the draft. Mark Boulos edited the draft. Yana Yunusova was responsible for conceptualization and editing the draft.

Data Availability Statement

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1 Croxson G, May M, Mester SJ. Grading facial nerve function: House-Brackmann versus Burres-Fisch methods. Am J Otol. 1990;11(4):240–6.
2 Enderby P. Frenchay dysarthria assessment. Int J Lang Commun Disord. 1980;15(3):165–73.
3 Scheller C, Wienke A, Tatagiba M, Gharabaghi A, Ramina KF, Scheller K, et al. Interobserver variability of the House-Brackmann facial nerve grading system for the analysis of a randomized multi-center phase III trial. Acta Neurochirurgica. 2017;159(4):733–8.

4 Rong P, Yunusova Y, Green JR. Speech intelligibility decline in individuals with fast and slow rates of ALS progression. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH; 2015 Jan; 2015. p. 2967–71.
5 Rong P, Yunusova Y, Richburg B, Green JR. Automatic extraction of abnormal lip movement features from the alternating motion rate task in amyotrophic lateral sclerosis. Int J Speech Lang Pathol. 2018;20(6):610–23.
6 Bandini A, Green JR, Wang J, Campbell TF, Zinman L, Yunusova Y, et al. Kinematic features of jaw and lips distinguish symptomatic from presymptomatic stages of bulbar decline in amyotrophic lateral sclerosis. J Speech Lang Hear Res. 2018;61(5):1118–29.
7 Chu SY, Barlow SM, Lee J. Face-referenced measurement of perioral stiffness and speech kinematics in Parkinson's disease. J Speech Lang Hear Res. 2015;58(2):201–12.
8 Bartle-Meyer CJ, Goozée JV, Murdoch BE, Green JR. Kinematic analysis of articulatory coupling in acquired apraxia of speech post-stroke. Brain Inj. 2009;23(2):133–45.
9 Kroos C. Evaluation of the measurement precision in three-dimensional electromagnetic articulography (Carstens AG500). J Phonetics. 2012;40(3):453–65.
10 States RA, Pappas E. Precision and repeatability of the Optotrak 3020 motion measurement system. J Med Eng Technology. 2006;30(1):11–6.
11 Savariaux C, Badin P, Samson A, Gerber S. A comparative study of the precision of carstens and northern digital instruments electromagnetic articulographs. J Speech Lang Hear Res. 2017;60(2):322–40.
12 Yunusova Y, Green JR, Mefferd A. Accuracy assessment for AG500, electromagnetic articulograph. J Speech Lang Hear Res. 2009;52(2):547–55.
13 Guarin DL, Yunusova Y, Taati B, Dusseldorp JR, Mohan S, Tavares J, et al. Toward an automatic system for computer-aided assessment in facial palsy. Facial Plast Surg Aesthet Med. 2020;22(1):42–9.
14 Guarin DL, Dempster A, Bandini A, Yunusova Y, Taati B. Estimation of orofacial kinematics in Parkinson's disease: comparison of 2D and 3D markerless systems for motion tracking. Conference: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020); 2020. p. 3–6.
15 Neumann M, Roesler O, Liscombe J, Kothare H, Suendermann-Oeft D, Pautler D, et al. Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale. 2021. Available from: http: //arxiv.org/abs/2104.07310.
16 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 Translational Neurol. 2020;7:1148–57.
17 Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). NPJ Digital Med. 2020;3(1):55.
18 FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) resource. 2021.
19 Ouni S, Dahmani S. Is markerless acquisition technique adequate for speech production? J Acoust Soc Am Acoust Soc America. 2016;139(6):EL234–9.
20 Guarin DL, Bandini A, Dempster A, Wang H, Rezaei S, Yunusova Y, et al. The effect of improving facial alignment accuracy on the video-based detection of neurological diseases. J Biomed Health Inform. 2020:1–9.
21 Robin J, Harrison JE, Kaufman LD, Rudzicz F, Simpson W, Yancheva M, et al. Evaluation of speech-based digital biomarkers: review and recommendations. Digit Biomark. 2020;4(3):99–108.
22 Vickers AJ. How many repeated measures in repeated measures designs? Statistical issues for comparative trials. BMC Med Res Methodol. 2003;3:22.
23 Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005;53(4):695–9.
24 Bulat A, Tzimiropoulos G. How far are we from solving the 2D and 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). Conference: International Conference on Computer Vision; 2017. p. 1021–30.
25 Murphy WK, Laskin DM. Intercanthal and interpupillary distance in the black population. Oral Surg Oral Med Oral Pathol. 1990;69(6):676–80.
26 Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334.
27 de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033–9.
28 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–8.
29 Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med. 2012;31(29):3972–81.
30 Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. J Am Stat Assoc. 1988;83(404):1014.
31 Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67(1).
32 Rossetti HC, Lacritz LH, Cullum CM, Weiner MF. Normative data for the Montreal Cognitive Assessment (MoCA) in a population-based sample. Neurology. 2011;77(13):1272–5.
33 Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). NPJ Digit Med. 2020;3.
34 Mokkink LB, Prinsen CAC, Bouter LM, Vet HC, Terwee CB. The COnsensus-based standards for the selection of health measurement INstruments (COSMIN) and how to select an outcome measurement instrument. Braz J Phys Ther. 2016;20(2):105–13.
35 Rong P, Yunusova Y, Eshghi M, Rowe HP, Green JR. A speech measure for early stratification of fast and slow progressors of bulbar amyotrophic lateral sclerosis: lip movement jitter. Amyotroph Lateral Scler Frontotemporal Degener. 2020;21(1–2):34–41.
36 Kuruvilla-Dugdale M, Mefferd A. Spatiotemporal movement variability in ALS: speaking rate effects on tongue, lower lip, and jaw motor control. J Commun Disord. 2017;67:22–34.
37 Mirelman A, Bernad-Elazari H, Thaler A, Giladi-Yacobi E, Gurevich T, Gana-Weisz M, et al. Arm swing as a potential new prodromal marker of Parkinson's disease. Mov Disord. 2016;31(10):1527–34.
38 Kay CD, Seidenberg M, Durgerian S, Nielson KA, Smith JC, Woodard JL, et al. Motor timing intraindividual variability in amnestic mild cognitive impairment and cognitively intact elders at genetic risk for Alzheimer's disease. J Clin Exp Neuropsychol. 2017;39(9):866–75.
39 Toosizadeh N, Ehsani H, Wendel C, Zamrini E, Connor KO, Mohler J, et al. Screening older adults for amnestic mild cognitive impairment and early-stage Alzheimer's disease using upper-extremity dual-tasking. Sci Rep. 2019;9:10911.
40 Asghari M, Ehsani H, Cohen A, Tax T, Mohler J, Toosizadeh N, et al. Nonlinear analysis of the movement variability structure can detect aging-related differences among cognitively healthy individuals. Hum Mov Sci. 2021;78:102807.
41 Aghanavesi S, Westin J, Bergquist F, Nyholm D, Askmark H, Aquilonius SM, et al. A multiple motion sensors index for motor state quantification in Parkinson's disease. Comput Methods Programs Biomed. 2020:105309.
42 Gudmundsson S, Runarsson TP, Sigurdsson S. Test-retest reliability and feature selection in physiological time series classification. Comput Methods Programs Biomed. 2012;105(1):50–60.
43 Yunusova Y, Green JR, Lindstrom MJ, Ball LJ, Pattee GL, Zinman L, et al. Kinematics of disease progression in bulbar ALS. J Commun Disord. 2010;43:6–20.

44 Rusz J, Tykalová T, Klempíř J, Čmejla R, Růžička E. Effects of dopaminergic replacement therapy on motor speech disorders in Parkinson's disease: longitudinal follow-up study on previously untreated patients. J Neural Transm. 2016;123(4):379–87.
45 Stegmann GM, Hahn S, Liss J, Shefner J, Rutkove SB, Kawabata K, et al. Repeatability of commonly used speech and language features for clinical applications. Digit Biomark. 2020;4(3):109–22.
46 Edwards LJ. Modern statistical techniques for the analysis of longitudinal data in biomedical research. Pediatr Pulmonol. 2000;30(4):330–44.
47 Vickers AJ. How many repeated measures in repeated measures designs? Statistical issues for comparative trials. 2003. Available from: http: //www.biomedcentral.com/1471-2288/3/22.