Relationship between sample size and responsiveness of speech-based digital biomarkers in ALS

Hardik Kothare[1], Michael Neumann[1] and Vikram Ramanarayanan[1,2]

[1] Modality.AI, Inc.,[2] University of California, San Francisco, CA, USA

hardik.kothare@modality.ai

Methodological Question and Introduction

- Clinical trials need **optimal sample size**, considering **budget constraints** and avoiding **underpowered trials**.
- **Costs** for assessing therapeutic benefits **increase exponentially** with more **patients** and **clinic visits**.
- **Smaller sample sizes** desirable for **ALS**, a rare neurodegenerative disorder with an estimated global prevalence of **4.42 per 100,000 people**.
- **Speech-based digital biomarkers** can **remotely track longitudinal progression** in people with Amyotrophic Lateral Sclerosis (pALS), i.e. **without clinic visits**. This study explores the **responsiveness of these biomarkers as a function of sample size**.

Data and Methods

	Number of participants	Number of sessions	Mean sessions per participant ± SD	Mean age ± SD (years)
Bulbar onset	36 (18 female)	598	16.6 ± 19.4	61.6 ± 11.9
Non-bulbar onset	107 (52 female)	2790	26.1 ± 25.9	59.9 ± 9.6

Table 1: Demographics

- Data collected using a **cloud-based multimodal dialogue platform** (Illustration in Figure 1)
- **Tina, a virtual guide,** walked participants through structured speaking exercises and **objective metrics** were extracted.
- Evaluation focused on the responsiveness of four **timing and intelligibility** related speech metrics calculated from read speech (**Bamboo passage, 99 words**) using Praat and the Montreal Forced Aligner:

Metric	Description
Speaking duration (s)	Time taken to read the reading passage.
Speaking rate (words per minute)	Number of words in the passage (99) divided by the time taken to read the reading passage.
Percentage pause time (PPT; %)	Total duration of all pauses divided by the total duration of the utterance expressed as a percentage.
Canonical Timing Alignment (CTA)	A number between 0% (non-alignment) and 100% (perfect alignment) as measured by the normalised inverse Levenshtein edit distance between words and silence boundaries. The participant's predicted word-level timing, obtained using the Montreal Forced Aligner, is compared to the expected production by Tina.

Table 2: Metrics

- **Growth curve models (GCMs,** Figure 2**)** used to estimate the trajectory of these metrics over time, with random slopes and intercepts for each participant.
- Responsiveness evaluated as: (i) time taken to detect deterioration greater than the standard error of the mean for the cohort (**statistical utility**) and (ii) time taken to detect deterioration greater than the minimal clinically-important difference (**clinical utility**) anchored to the ALS Functional Rating Scale - Revised (**ALSFRS-R**) scale.
- To investigate the relationship between responsiveness and sample size of the participant cohort, sample sizes of **30**, **25**, **20**, **15** and **10** participants were **randomly sampled 100 times**, without replacement, from both cohorts. GCMs were run for each of these 100 iterations.
- **Mean responsiveness** calculated as the average slope for each cohort across 100 iterations.

Acknowledgements

This work was supported by the National Institutes of Health grant R42DC019877. We thank our collaborators at EverythingALS and the Peter Cohen Foundation for participant recruitment and data collection. Disclosure: All authors are full-time employees of Modality.AI and hold stock options in the company.

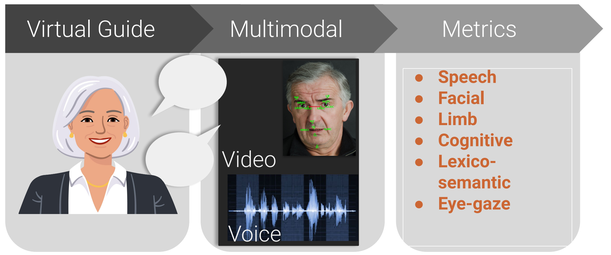

Figure 1. Schematic of the Modality.AI dialogue platform

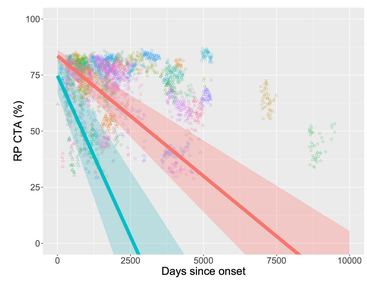

Figure 2. Example of a Growth Curve Model (GCM). RP = Reading Passage

Results and Discussion

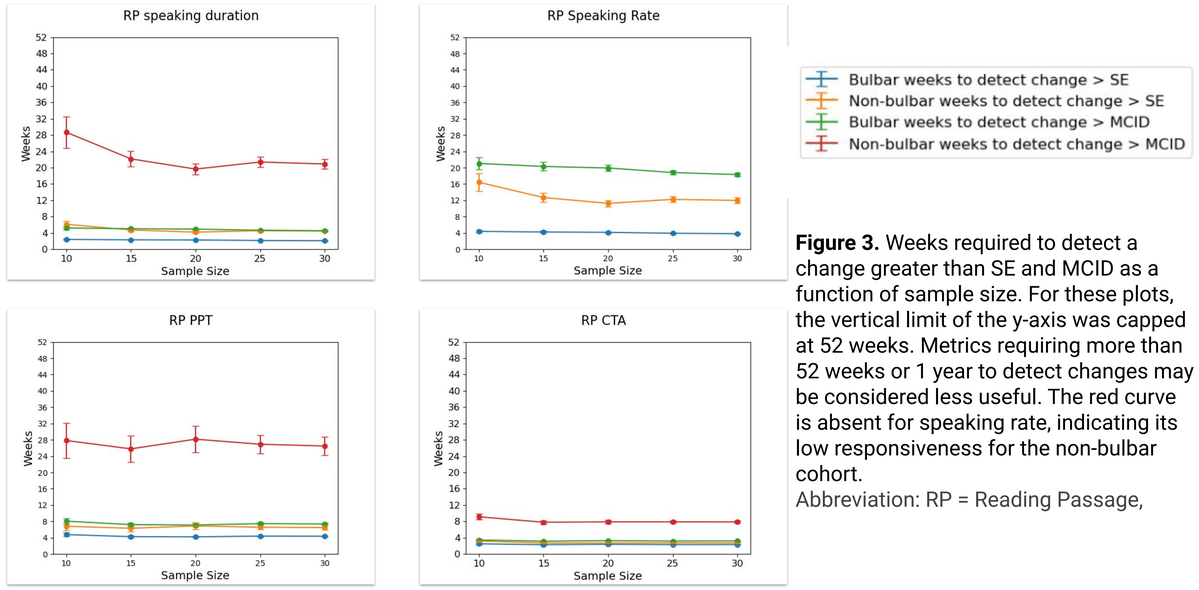

Figure 3. Weeks required to detect a change greater than SE and MCID as a function of sample size. For these plots, the vertical limit of the y-axis was capped at 52 weeks. Metrics requiring more than 52 weeks or 1 year to detect changes may be considered less useful. The red curve is absent for speaking rate, indicating its low responsiveness for the non-bulbar cohort.
Abbreviation: RP = Reading Passage,

- Mean responsiveness of the four biomarkers remains **stable** even with **15 people per cohort**.
- **Confidence interval** for mean responsiveness **increases** with **decreasing sample size**.
- For **non-bulbar** pALS, detecting a change > MCID in **speaking rate** takes **more than 52 weeks**.
- **CTA** is **highly responsive**, detecting clinically-important changes within **3.22** (± 0.07) to **3.46** (± 0.25) weeks in the bulbar cohort and within **7.88** (± 0.24) to **9.14** (± 0.68) weeks in the non-bulbar cohort, as the sample size decreases from 30 to 10.

Conclusions

- **Speech-based digital biomarkers** show promise in enabling ALS clinical trials with **small sample sizes**.
- The **relationship** between **sample size** and **mean responsiveness** is **stable** when sample sizes range between 10 and 30 participants per cohort, but **uncertainty increases with smaller sizes**, necessitating consideration in clinical trial design.

References

- Xu, Lu, et al. "Global variation in prevalence and incidence of amyotrophic lateral sclerosis: a systematic review and meta-analysis." Journal of Neurology 267 (2020): 944-953
- Ramanarayanan, V., et al. "When Words Speak Just as Loudly as Actions: Virtual Agent Based Remote Health Assessment Integrating What Patients Say with What They Do." Proc. INTERSPEECH (2023), 678-679
- Kothare, H., et al. "Responsiveness, Sensitivity and Clinical Utility of Timing-Related Speech Biomarkers for Remote Monitoring of ALS Disease Progression." Proc. INTERSPEECH (2023), 2323-2327