Development and validation of speech-based biomarkers for measuring clinical progression in AD clinical trials

WINTERLIGHT

CAMBRIDGE COGNITION

Michael Spilka[1], Mengdan Xu[1], Bali Toth[2], Somaye Hashemifar[2], Rainier Amora[2], Jessica Robin[1], Edmond Teng[2], Cecilia Monteiro[2], & William Simpson[1]

1. Winterlight Labs, Inc. (a division of Cambridge Cognition), Toronto, ON, Canada. 2. Genentech, Inc., South San Francisco, CA, USA

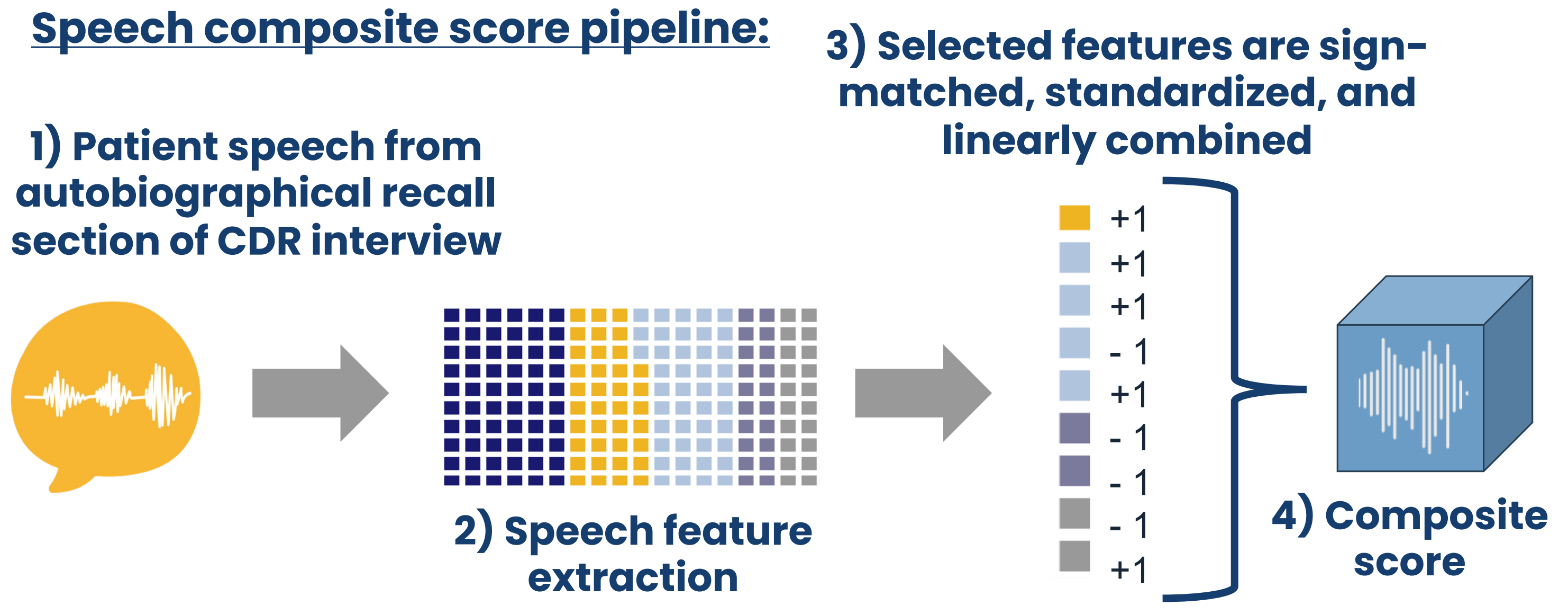
Contact: michael.spilka@camcog.com

Background

- Progressive language changes are established clinical characteristics of Alzheimer's disease (AD).
- Advances in Natural Language Processing (NLP) enable more objective, nuanced measurement of language, facilitating the development of speech biomarkers for tracking longitudinal decline in language function.
- **Objective:** We evaluated and compared several low-burden, digital speech-based markers developed from clinical interview recordings from two phase 2 clinical trials.

Methods

- Participants: 227 English-speaking individuals pooled from two phase 2 trials of semorinemab: Tauriel (MCI-to-mild AD; [NCT03289143]) and Lauriet (mild-to-moderate AD; [NCT03828747]).
- Clinical Dementia Rating (CDR) interview recordings were analyzed at screening, baseline, week 25, and week 49, focusing on participant speech from the autobiographical recall section of the interview.
- Data were split 60%/40% into training and testing sets for development and validation of speech composite scores.
- Three speech feature selection approaches were evaluated:
  1. **Replication composite:** features from our previously published 9-feature AD speech composite score (Robin et al., 2023; *Alzheimer's & Dementia: DADM*. doi: 10.1002/dad2.12445).
  2. **Novel composite 1:** features with a stringent $p < .001$ effect of change over time (12 features).
  3. **Novel composite 2:** features with a $p < .05$ effect of time, ICC > 0.5, and prioritizing clinical interpretability (e.g., linguistic vs. signal-processing features; 18 features).
- Speech composites were evaluated on:
  1. Longitudinal change (time effect from linear mixed models adjusting for age, gender, education).
  2. Test-retest reliability (screening vs. baseline visit intraclass correlations; ICCs).
  3. Correlations with clinical endpoints (Spearman correlations with ADAS-Cog11, CDR-SB, ADCS-ADL, MMSE).


Speech composite score pipeline:

1) Patient speech from autobiographical recall section of CDR interview

2) Speech feature extraction

3) Selected features are sign-matched, standardized, and linearly combined

4) Composite score

Participant characteristics

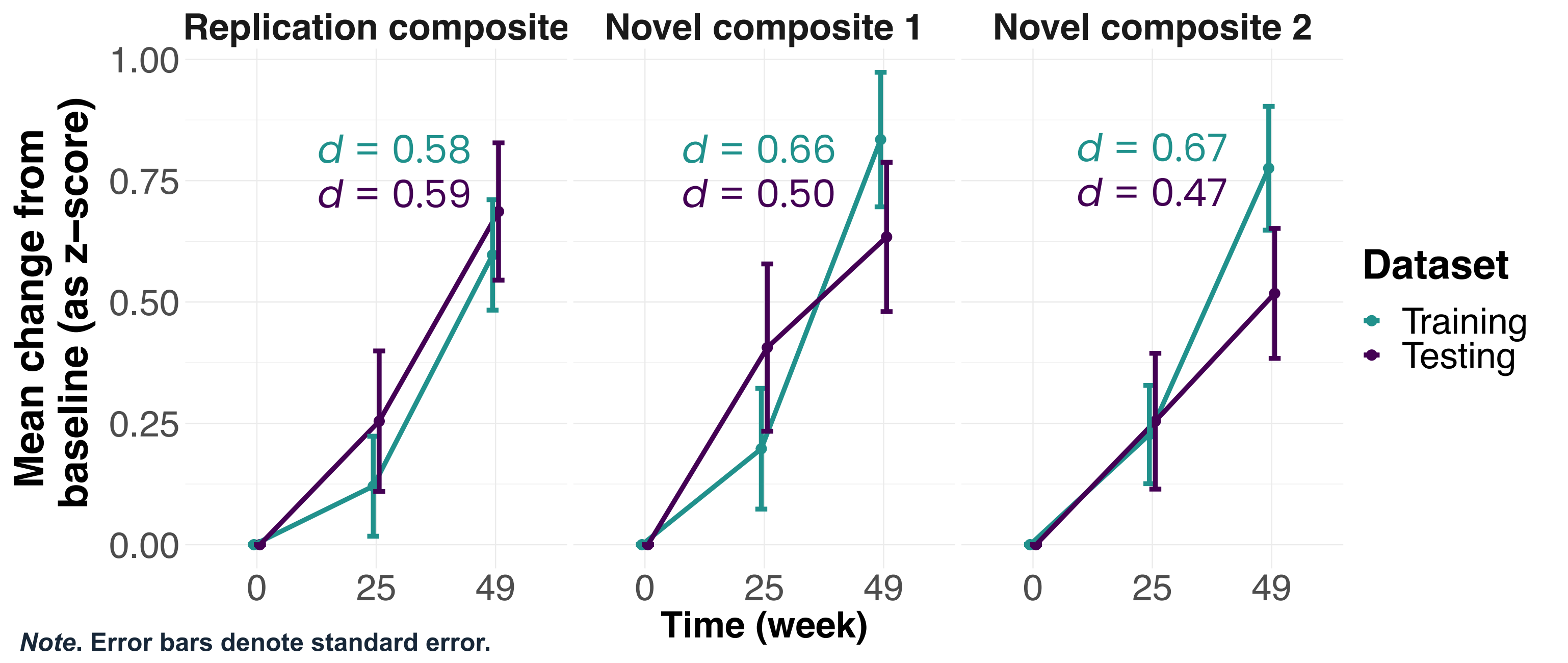
- The training and testing datasets did not significantly differ on clinical scores at baseline or their longitudinal trajectories.

Baseline characteristics	Training (60%)	Testing (40%)	*p*-value
n	Tauriel: 87 Lauriet: 48	Tauriel: 60 Lauriet: 32	1
Age (*M*, *SD*)	70.3 (8.4)	70.6 (7.8)	.79
Sex (n, %)			.50
Female	77 (57%)	57 (62%)	
Male	58 (43%)	35 (38%)	
ADAS-Cog11 Total (*M*, *SD*)	19.9 (6.6)	19.5 (7.2)	.62
CDR-SB (*M*, *SD*)	4.8 (2.1)	4.7 (2.1)	.72
ADCS-ADL Total (*M*, *SD*)	66.6 (7.7)	65.8 (9.1)	.47
MMSE Total (*M*, *SD*)	21.5 (3.5)	21.5 (3.7)	.89

Note. MMSE = Mini Mental State Examination. ADAS-Cog11 = Alzheimer's Disease Assessment Scale–Cognitive Subscale. CDR-SB = Clinical Dementia Rating Scale Sum of Boxes. ADCS-ADL = Alzheimer's Disease Cooperative Study – Activities of Daily Living Scale.

Results: Longitudinal change

- All 3 composites showed significant change over time (training set: β = 0.51-0.68; testing set: β = 0.49-0.61; *p*'s < .001), with medium effect sizes of baseline to endpoint change scores (Cohen's *d*).



Note. Error bars denote standard error.

Results: Test-retest reliability

- Intraclass correlations (ICCs) for Screening vs. Baseline scores indicated moderate-to-good reliability for all 3 composites, with the highest in the testing set for the replication composite (ICC = 0.80).

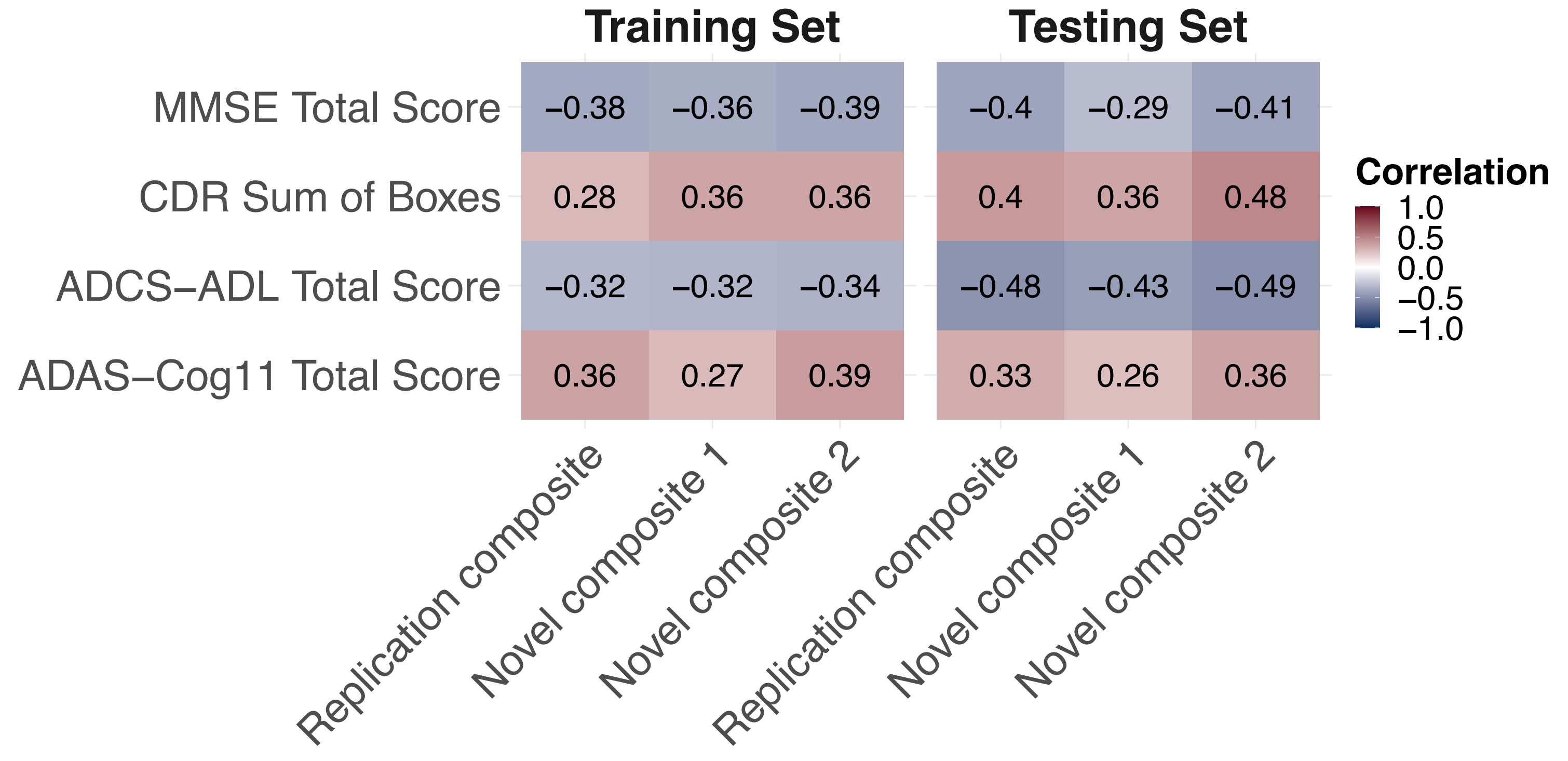
Composite score	Screening vs. Baseline test-retest reliability (ICC)	
	Training set	Testing set
Replication composite	0.73	0.80
Novel composite 1	0.59	0.67
Novel composite 2	0.77	0.76

9-feature speech composite biomarker of clinical progression in AD

Word length	Noun use	MFCC 11 mean	Linguistic Acoustic
Syntactic depth	Particle use	MFCC 25 variance	
Word frequency	Pronoun use	MFCC 28 variance	

Results: Baseline correlations with clinical endpoints

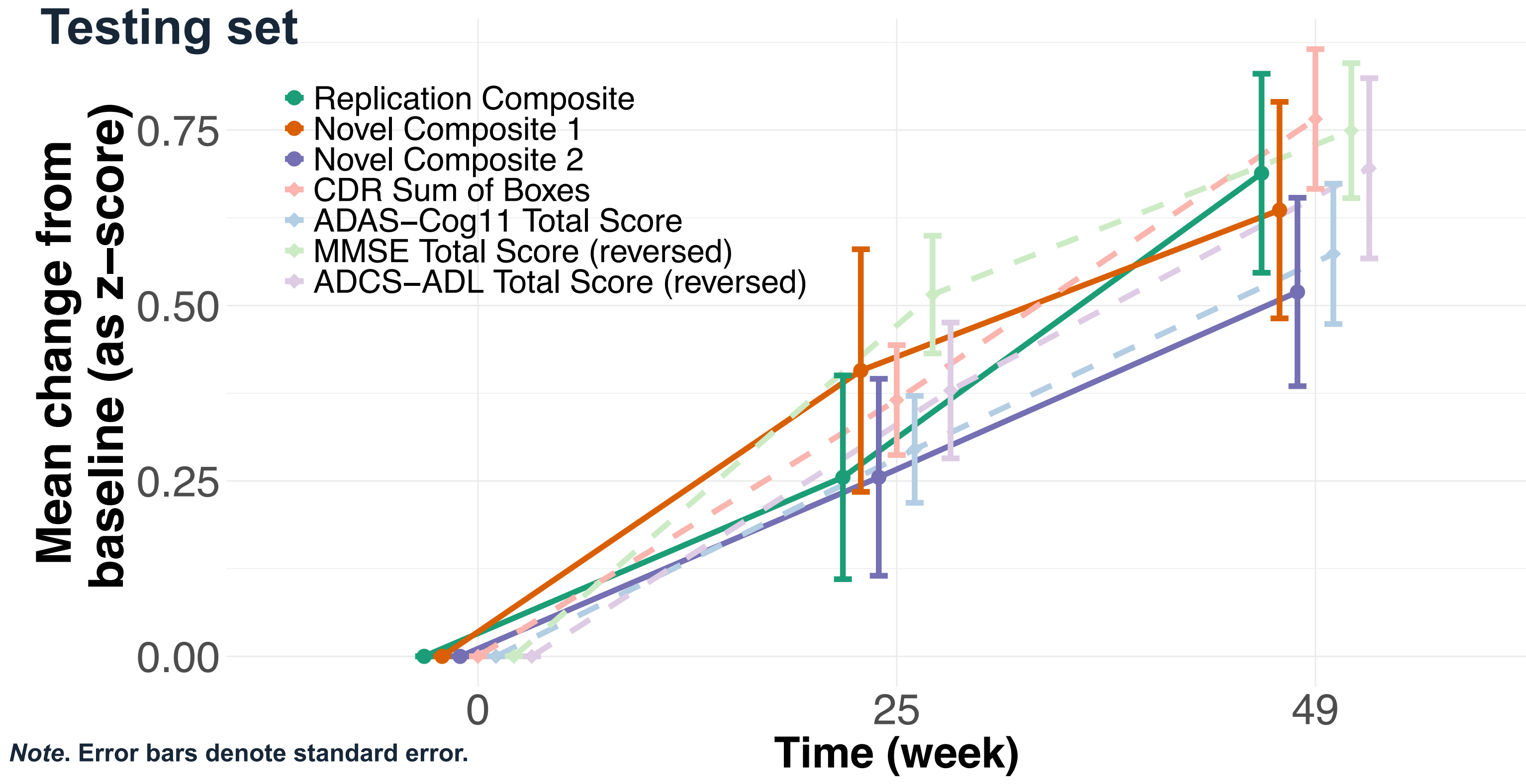
- At baseline, all three speech composites were significantly correlated with the study clinical endpoints (small-to-moderate correlation strength).



	Training Set: Replication composite	Training Set: Novel composite 1	Training Set: Novel composite 2	Testing Set: Replication composite	Testing Set: Novel composite 1	Testing Set: Novel composite 2
MMSE Total Score	−0.38	−0.36	−0.39	−0.4	−0.29	−0.41
CDR Sum of Boxes	0.28	0.36	0.36	0.4	0.36	0.48
ADCS–ADL Total Score	−0.32	−0.32	−0.34	−0.48	−0.43	−0.49
ADAS–Cog11 Total Score	0.36	0.27	0.39	0.33	0.26	0.36

Results: Longitudinal comparisons with clinical endpoints

- Speech composite scores generally demonstrated similar sensitivity to clinical progression (testing set: *d* = 0.47-0.59) as the study efficacy endpoints (*d* = 0.59-0.85).



Note. Error bars denote standard error.

Conclusions

- Each speech composite score performed well overall. **The best performing composite was our previously published Tauriel-derived speech biomarker**: it had the largest effect size of change, highest test-retest reliability, and was the most parsimonious measure with the fewest features.
- These results highlight the potential utility of a **speech-based biomarker** as an **objective** and **low- burden measure of clinical progression** to complement traditional endpoints in **AD clinical trials**.