Benchmarking Prognostic Longitudinal Machine Learning Models of Alzheimer's Disease Using Speech Features

Malikeh Ehghaghi[1], Jekaterina Novikova[2], Arindam Sett[3], Mohsen Hejrati[4], Jessica Robin[5] , Edmond Teng[6], Somaye Hashemifar[7]

[1,2,5] Winterlight Labs Inc., Toronto, Canada

[3,4,6,7]Genentech Inc., South San Francisco, CA, United States

{malikeh[1], jekaterina[2], jessica[5]}@winterlightlabs.com {setta[3], hejratis[4], tenge1[6], hashems4[7]}@gene.com

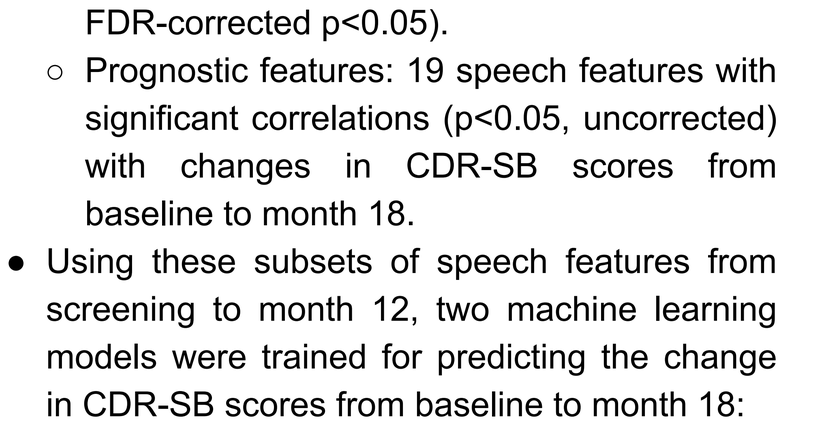
Background

The potential applicability of speech-related biomarkers for Alzheimer's disease (AD) in prognosis of longitudinal outcomes is largely unknown. Most research is focused on diagnostic classification models[1,2] or predicting Mini-Mental State Examination (MMSE) scores[3]. We developed two predictive models of clinical progression in prodromal-to-mild AD [measured by the Clinical Dementia Rating-Sum of Boxes (CDR-SB)[4]] that use a combination of several linguistic and acoustic speech features.

Methods

- Speech and CDR-SB data from a subset of 54 participants with prodromal-to-mild AD from the Tauriel study[5] of the anti-tau antibody semorinemab (NCT03289143) were analyzed.
- Data were obtained at five different visits (screening, baseline, months 6, 12, and 18) and pooled across semorinemab and placebo arms given the similar rates of clinical progression.
- Using the Winterlight speech processing pipeline, 520 acoustic and linguistic features were extracted from the CDR speech recordings and subsets of cross-sectional, and prognostic features were identified:
  - Cross-sectional features: 17 speech features with the strongest Pearson correlations with CDR-SB scores at baseline ($r>0.3$, FDR-corrected $p<0.05$).
  - Prognostic features: 19 speech features with significant correlations ($p<0.05$, uncorrected) with changes in CDR-SB scores from baseline to month 18.
- Using these subsets of speech features from screening to month 12, two machine learning models were trained for predicting the change in CDR-SB scores from baseline to month 18:
  1) Mixed Effects Random Forest (MERF)[6]
  2) Long Short-Term Memory (LSTM)[7]
- The performance of the models were evaluated based on mean absolute error (MAE), and root mean-squared error (RMSE) for both 5-fold cross-validation and held-out test sets.
  Note! To investigate the strength of our models, the performance of no change baseline model was reported that always predicts 0 changes in scores from baseline to month 18 sessions.

Figure 1: Winterlight Speech Analysis Pipeline



Results

- With either feature subset, MERF and LSTM models achieved less than 2.7 points of MAE for absolute change in CDR-SB scores (range: 0-18; Table 1 and 2).
- The best-performing MERF model outperformed the LSTM and no change baseline models on both train and held-out test sets when using cross-sectional features.
- MERF with cross-sectional features performed better compared to the prognostic features on both training and test data sets.

Figure 2: True vs. predicted CDR-SB change in MERF with cross-sectional features.

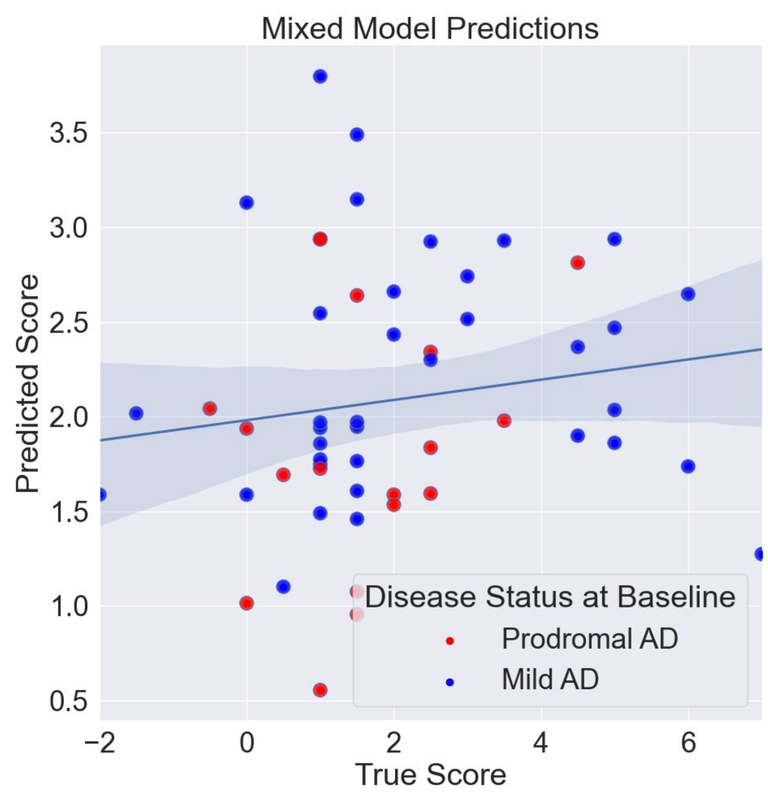


Table 1: Performance of the models predicting CDR-SB change from baseline to month 18 using 5-fold cross-validation on the train set

Predictive Model	MAE	RMSE
No change baseline	2.19±0.000	2.77±0.000
MERF with prognostic	1.62±0.038	2.06±0.031
MERF with cross-sectional	**1.48±0.040**	**1.95±0.059**
LSTM with prognostic	1.59±0.076	2.10±0.083
LSTM with cross-sectional	1.62±0.059	2.18±0.074

Table 2: Performance of the models predicting CDR-SB change from baseline to month 18 on the held-out test set

Predictive Model	MAE	RMSE
No change baseline	2.59	3.50
MERF with cross-sectional	**2.10**	**2.80**
LSTM with cross-sectional	2.67	3.26

Conclusion

- We developed several longitudinal models for predicting clinical progression in AD using speech features.
- Our results suggest that the nonlinear mixed effect model is efficacious in longitudinal monitoring of AD.
- Our results also signify that cross-sectional speech features, which are significantly correlated with CDR-SB scores at baseline assessment, can be used as effective predictors of cognitive decline across 18 months of follow-up.

References

(1) Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge. arXiv preprint arXiv:2004.06833.

(2) Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2021). Detecting cognitive decline using speech only: The ADReSSo Challenge. arXiv preprint arXiv:2104.09356.

(3) Yancheva, M., Fraser, K. C., & Rudzicz, F. (2015, September). Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies* (pp. 134-139).

(4) O'Bryant, S. E., Waring, S. C., Cullum, C. M., Hall, J., Lacritz, L., Massman, P. J., ... & Texas Alzheimer's Research Consortium. (2008). Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores: a Texas Alzheimer's research consortium study. Archives of neurology, 65(8), 1091-1095.

(5) Teng, E., Manser, P. T., Pickthorn, K., Brunstein, F., Blendstrup, M., Bohorquez, S. S., ... & Tauriel Investigators. (2022). Safety and efficacy of semorinemab in individuals with prodromal to mild Alzheimer disease: a randomized clinical trial. JAMA neurology, 79(8), 758-767.

(6) Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation, 84(6), 1313-1328.

(7) Graves, A. (2012). Long short-term memory. Supervised sequence labelling with recurrent neural networks, 37-45.

Presented at Applications of Medical AI (AMAI), 2022