University of Zurich[UZH]

EBPI Epidemiology, Biostatistics and Prevention Institute

RandomForests4Life: A Statistical Approach to Modelling ALS Progression

Torsten Hothorn

2014-07-01

ALS

- Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease.
- Substantial heterogeneity leading to difficulties in clinical practice as well as in estimating the effectiveness of new treatments.
- Better tools for determining disease progression needed.
- We introduce Prize4Life, PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials Database), Prize4Life and RandomForest4Life.

Prize4Life

- Prize4Life is a nonprofit organization dedicated to accelerating the discovery of treatments and cures for ALS
- It was founded by a group of Harvard Business School students when one of them, Avichai Kremer, was diagnosed with ALS.
- The group wants to accelerate ALS research by offering substantial prizes to scientists who solve the most critical scientific problems preventing the discovery of an effective ALS treatment.

http://www.prize4life.org

The Problem

- Identification of prognostic factors and corresponding models for predicting the disease progression in ALS is a long-standing problem.
- Having such an instrument would allow the planning of more powerful clinical trials by means of efficient patient stratification.
- Approaches exist for predicting overall survival and function via the ALSFRS (ALS functional rating scale) score.
- Published prognostic factors are bulbar rather than limb onset, BMI, early disease progression, age at onset, uric acid level, and amount of repeat expansion in gene C9ORF72.

The Challenge

- To stimulate collaborative research efforts for the development of prediction models the DREAM project (Dialogue for Reverse Engineering Assessments and Methods, sponsored by IBM, Columbia University, NIH Roadmap Initiative, and The New York Academy of Sciences) and Prize4Life jointly launched the DREAM Phil Bowen ALS Prediction Prize4Life Challenge on 10 July 2012.
- The challenge asked for submissions describing a prediction model for ALS disease progression.
- 1073 people registered for the challenge.
- 37 teams submitted a proposal.
- Three teams were awarded a prize of (in total) USD 50,000.

https://nctu.partners.org/ProACT/

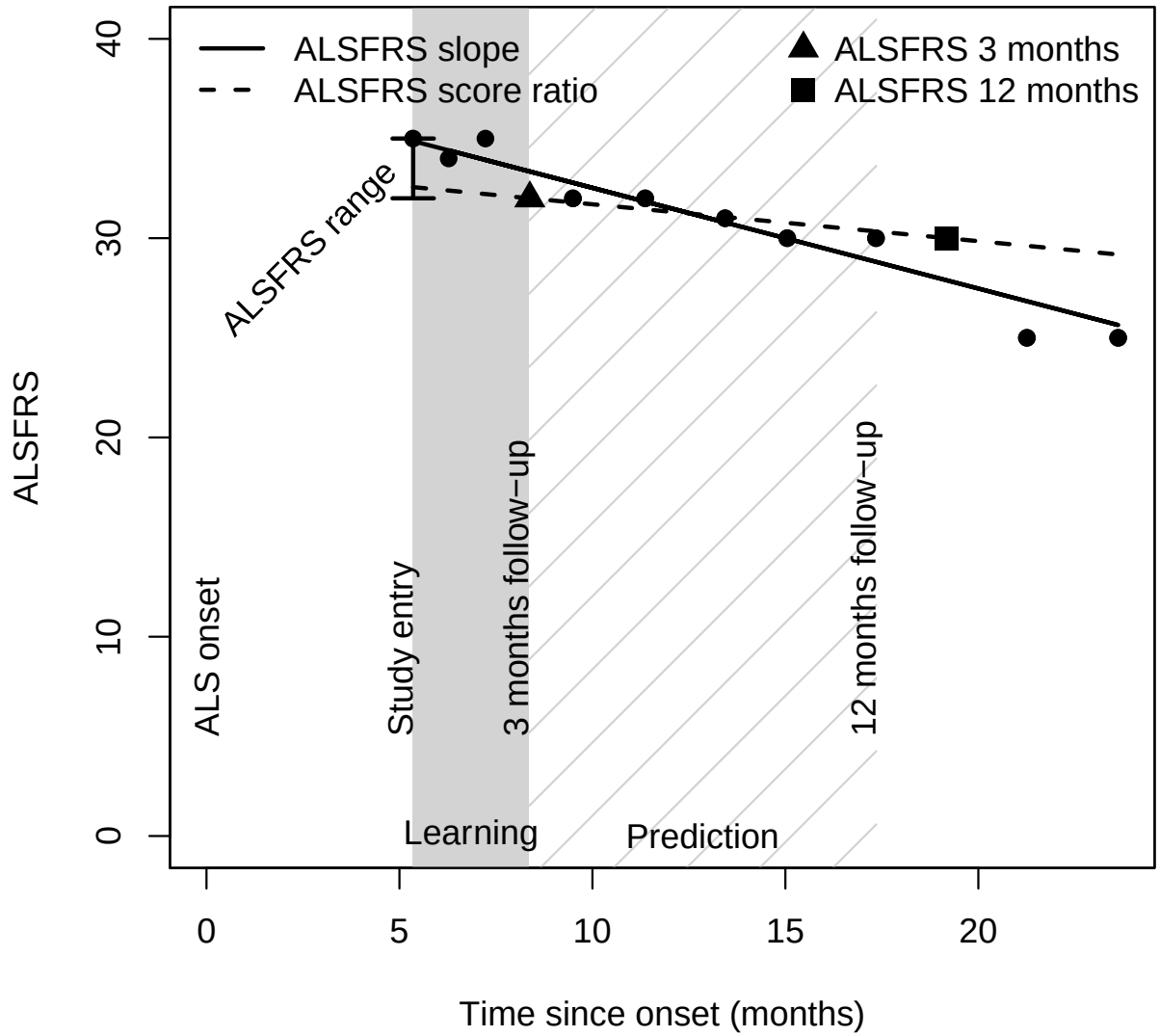
PRO-ACT

- Participants were provided with access to standardized, anonymized Phase II/III clinical trial data from 1822 ALS patients.
- Patient data came from the PRO-ACT database, now containing clinical trial data from 8,500 ALS patients from multiple trials.
- Patient information included demographics, family history, medical history, physiological and laboratory parameters, and ALSFRS readings.
- Patient survival was not available at the time but is now.

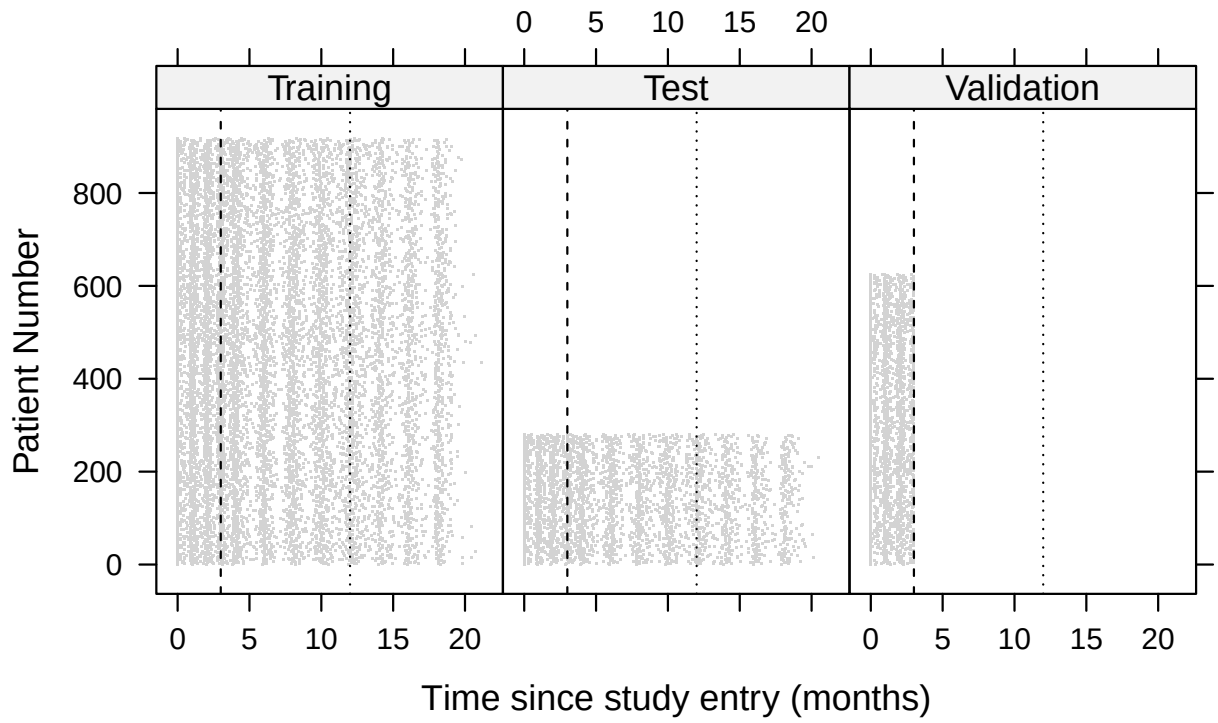
The Task

- Disease progression was measured by ALSFRS trajectories over time.
- Methods were sought for predicting the standardized difference between the ALSFRS readings taken approximately at three and twelve months after study entry for 625 patients in a "validation sample" based on information obtained in the first three months.
- Another set of 1197 with full information over time was available for "learning" such a prediction model.

Schedule Overview



Study Overview



The ALSFRS Score Ratio

- The standardized difference between two ALSFRS readings is called "score ratio" (Kollewe et al., 2008).
- It implies that the ALSFRS trajectory changes linearily between three and twelve months.
- So, in fact we are interested in the slope of a linear function fitted to the ALSFRS trajectories for each patient.
- But we do have information on much more ALSFRS readings for each patient, not just two!
- How can we use this information to obtain a better measure of disease progression?

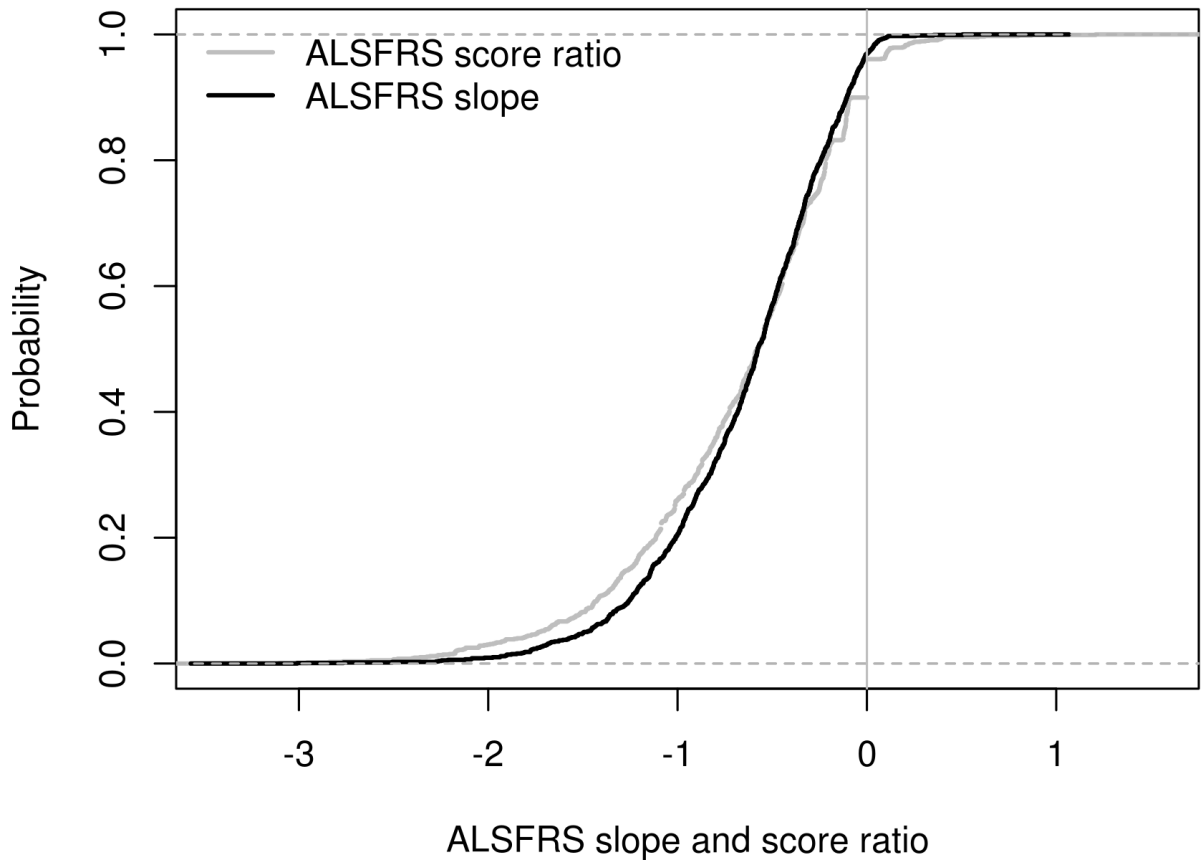
The ALSFRS Slope

- Idea: Fit a linear function to the ALSFRS trajectory of each patient and use it's slope as measure of disease progression.
- $A_{it} \in \{0, \ldots, 40\}$ denotes the ALSFRS score of patient i read at some time t after disease onset.
- Model:

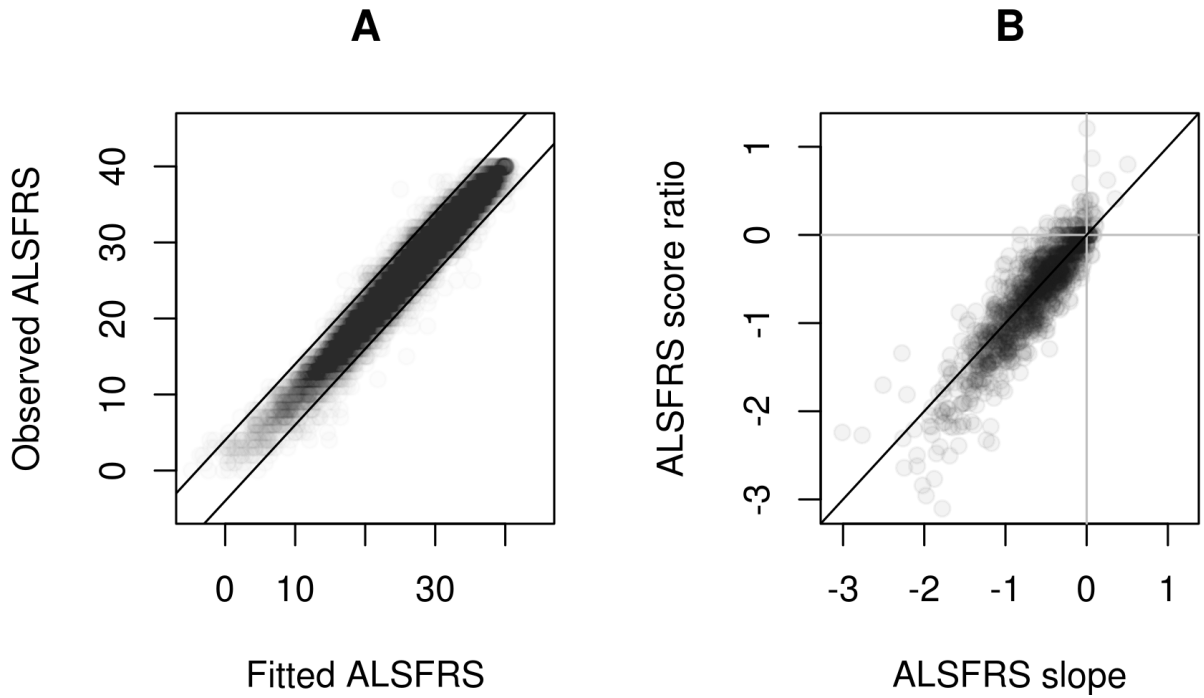
$$\mathrm{E}(A_{it}) = 40 + \alpha_i + (\beta_1 + \gamma_i)t$$

- γ_i can be interpreted as the deviation of the slope for patient i from the mean slope β_1 that applies to all patients.
- This is a linear mixed model using ALL the ALSFRS data from each patient.
- We use $\hat{\beta}_1 + \hat{\gamma}_i$ as new target variable for prediction.

ALSFRS Score Ratio vs. ALSFRS Slope



ALSFRS Score Ratio vs. ALSFRS Slope



Random Forests

- A Random Forest is a statistical learning approach and one of the best "of-the-shelf" prediction methods.
- It basically works by splitting the patients into multiple homogeneous subgroups wrt ALSFRS slopes.
- Subgroups are described by patient characteristics.
- We do this over and over again after resampling the patients.
- Finally, we average the ALSFRS slopes of "similar" patients, i.e., patients which ended up in the same subgroup often.
- This average serves as the prediction.
- Sounds simple but is in fact a little tricky when it comes to the details (but that's my job!).

Baseline Variables

Demographics: age, sex, race, height, affected region at onset (onset site), and time since onset.

Family history: family members affected by ALS (all binary variables): aunt, aunt (maternal), cousin, father, grandfather, grandfather (maternal), grandfather (paternal), grandmother, grandmother (maternal), grandmother (paternal), mother, niece, uncle, uncle (maternal), uncle (paternal), son, daughter, sister, brother.

Medical history: previously diagnosed neurological diseases (all binary variables): atrophy, cramps, fasciculations, gait changes, sensory changes, stiffness, speech, swallowing, weakness, others.

Patient Characteristics

		Training	Test	Validation
Sex				
	Female	53	182	107
	Male	91	303	222
	na	135	433	296
Race				
	Caucasian	0	8	5
	Asian	4	14	9
	Black/African Am.	272	886	598
	na	3	10	13
Onset site				
	Bulbar	59	203	121
	Limb	218	711	500
	Limb and bulbar	2	4	4
Age (years)		53 (44–62)	55 (47–64)	56 (47–64)
Height (cm)		170 (164–178)	170 (163–178)	172 (165–179)
Weight (kg)		74 (63–86)	73 (60–85)	71 (60–84)
Number of visits		12 (11–12)	12 (11–12)	4 (3–4)
3-12 month period (days)		277 (258–305)	279 (260–303)	na

Time-varying Variables (I)

ALSFRS(-R): all ALSFRS and ALSFRS-R items, i.e., speech, salivation, swallowing, handwriting, cutting, dressing and hygiene, turning in bed, walking, climbing stairs, respiratory (ALSFRS only), dyspnea (ALSFRS-R only), orthopnea (ALSFRS-R only), respiratory insufficiency (ALSFRS-R only) and the corresponding sum scores. In addition, we used the range of the ALSFRS score in the first three months as measure of variability.

Physiological parameters: patient weight, blood pressure (systolic and diastolic), pulse rate, respiratory rate, slow and forced vital capacity.

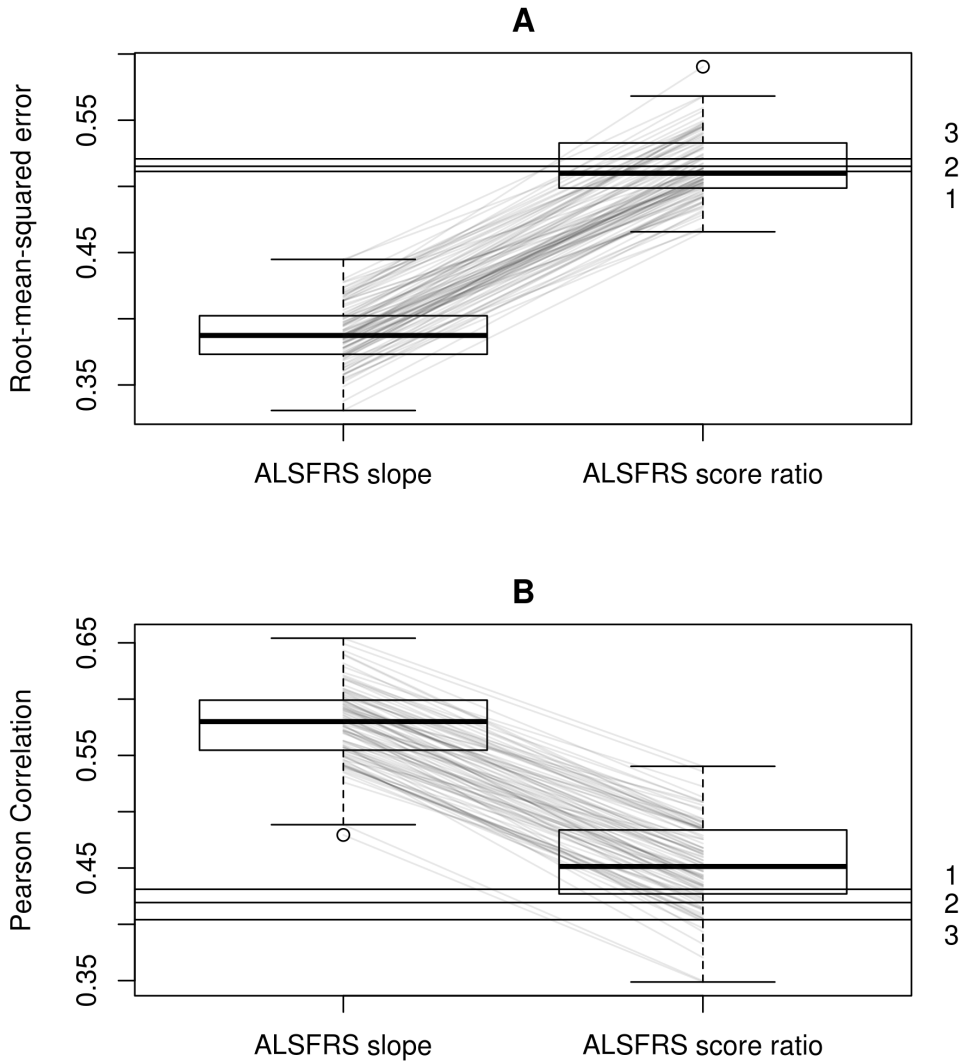
Time-varying Variables (II)

Laboratory parameters: alkaline phosphatase (11% mean proportion of missing values over all patients), chloride (20%), creatinine (11%), ASTSGOT (11%), neutrophils (12%), protein (11%), calcium (11%), glucose (11%), blood urea nitrogen (11%), bicarbonate (26%), bilirubin total (11%), phosphorus (11%), ALTSGPT (11%), triglycerides (20%), hematocrit (12%), creatine kinase (20%), eosinophils (12%), lymphocytes (12%), albumin (11%), white blood cells (18%), red blood cells (18%), absolute basophil count (86%), HbA1c glycated hemoglobin (26%), platelets (12%), total cholesterol (11%), sodium (11%), monocytes (12%), gamma glutamyltransferase (11%), hemoglobin (12%), potassium (11%), basophils (12%), urine glucose (87%), urine protein (87%), urine pH (15%).

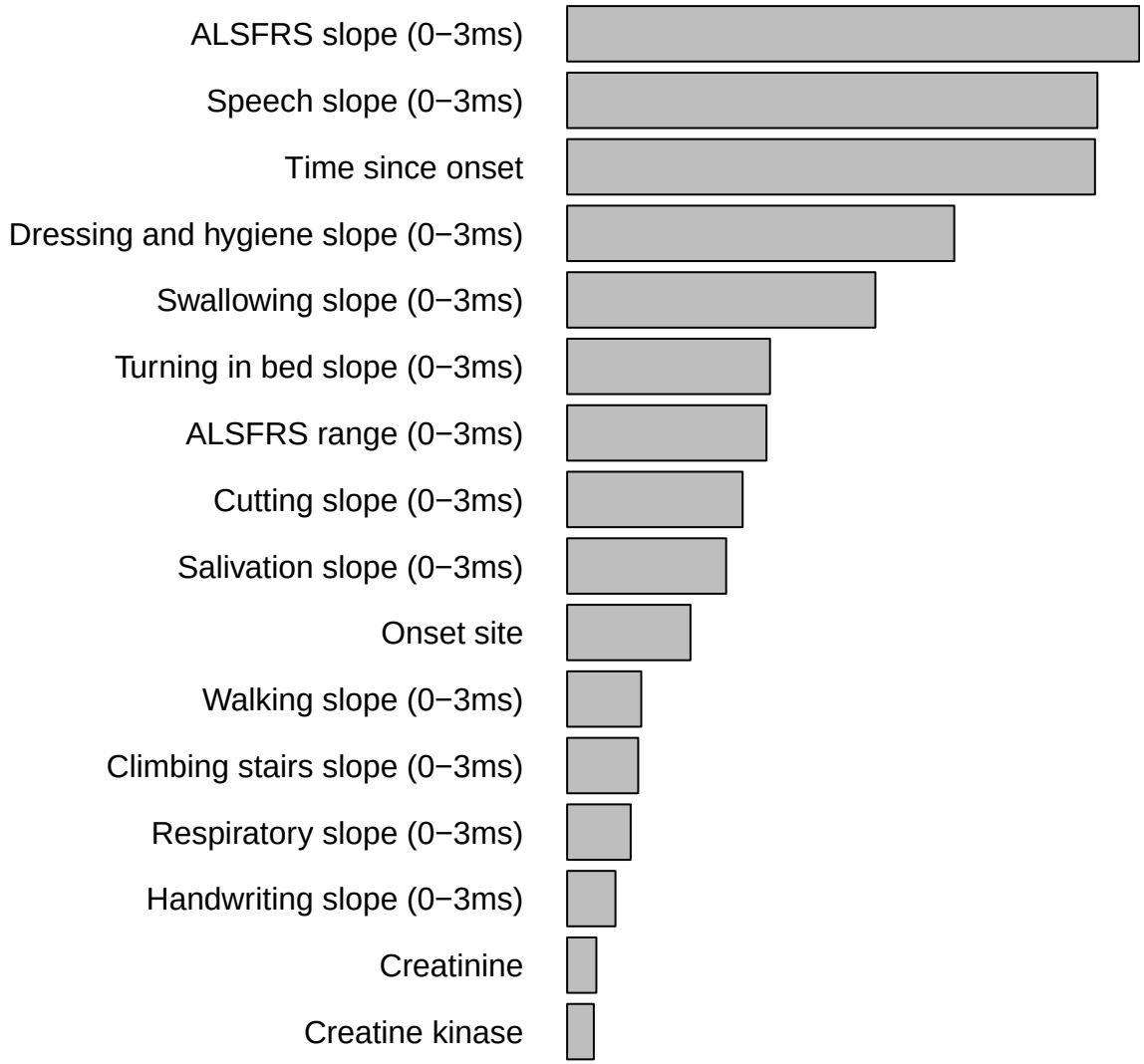
Results: Prediction

- The submissions were reviewed by an expert panel.
- The quality of the prediction was measured by the root mean squared error between the predicted and the observed ALSFRS score ratio (!) for the 625 patients in the validation sample.
- Also, the correlation was looked at.
- Two teams secured first prize, a duo from Stanford University, Lester Mackey, PhD, and Master's Degree recipient Lilly Fang; and the team of Liuxia Want, PhD, and her colleague Guang Li, Quantitative Modeler the scientific marketing company Sentrana.
- Our approach was awarded a second-place prize.

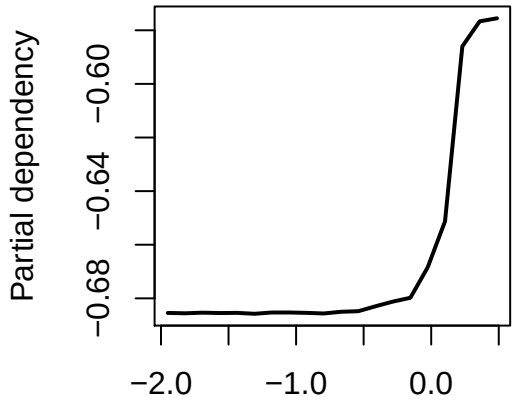
Results: Prediction

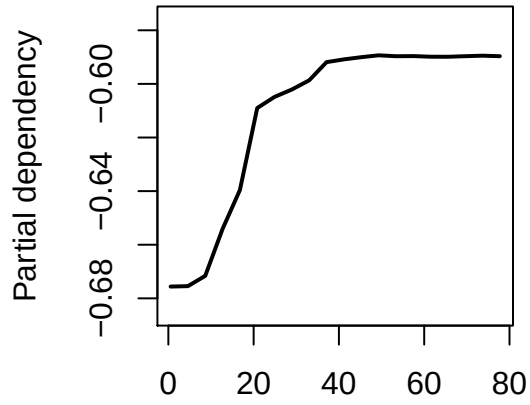


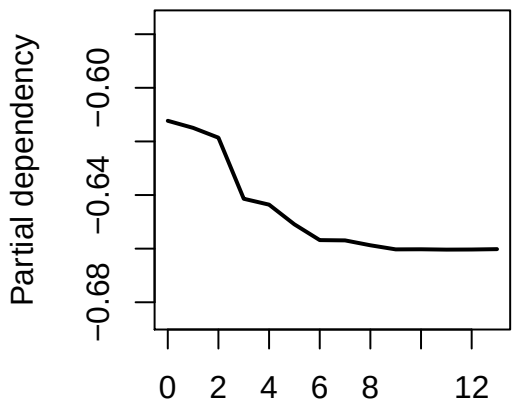
Results: Important Variables

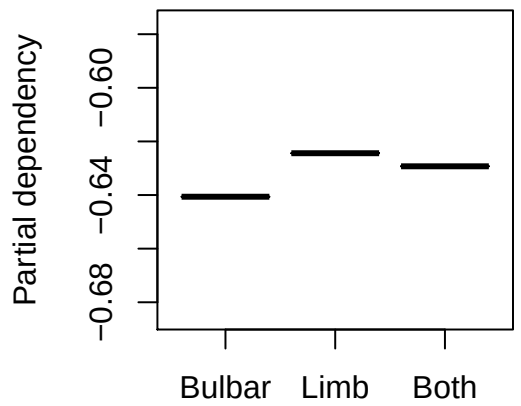


Results: Important Variables









Summary

- The DREAM-Phil Bowen ALS Prediction Prize4Life was the first crowdsourcing challenge to use clinical trial data.
- The best performing algorithms outperformed a method designed by the challenge organizers as well as predictions by ALS clinicians.
- Using these methods may reduce the sample size of future clinical trials through efficient patient stratification by 20%.
- The results suggest several novel predictors.
- These and other details in Hothorn & Jung (ALSFD, 2014) and Küffner et al. (NBT under revision).

References

Hothorn & Jung (2014)
http://dx.doi.org/10.3109/21678421.2014.893361

Kollewe et al. (2008)
http://dx.doi.org/10.1016/j.jns.2008.07.016