Discovering New Variables and Improving the Prediction of ALS Progression

Guang Li, Liuxia Wang

Sentrana Inc.

November 13, 2012

RECOMB and DREAM Challenge, San Francisco

Agenda

1. Problem Description, Challenges and Goals
2. Data Preparation
3. Model Specification
4. Analysis and Results
5. Interesting Implications
6. Appendix

Problem Description, Challenges and Goals

SENTRANA

1. Problem Description:

1) Problem:

- Different ALS progression rate among patients
- Use 3 month data to predict next 9 months progression rate

2) Measurement of ALS progression:

- ALS Functional Rating Scale (ALSFRS): 10 questions, each question scale 0-4, totally 0-40
- High correlation of ALSFRS score change with patients' survival time
- Response variable: (ALSFRS score at last trial – ALSFRS score at first trial)/Time difference

3) Data:

- PRO-ACT (Pooled Resource Open-access ALS Clinical Trials) , 1197 patients
- Data includes: ALSFRS historical data, family history, demographic data, onset/diagnosis data, vital capacity, vital signs and several different lab tests etc.

2. Challenges:

1) Unknown causality of ALS
2) Dirty data
3) Many input variables

3. Goals:

1) Discover more predictors
2) Design algorithm with better accuracy

Data Preparation: Exploratory Analysis

1. Choice of variables and variable definition are based on both the literature review and exploratory analysis.

1) Factors found in literature review are summarized as below:

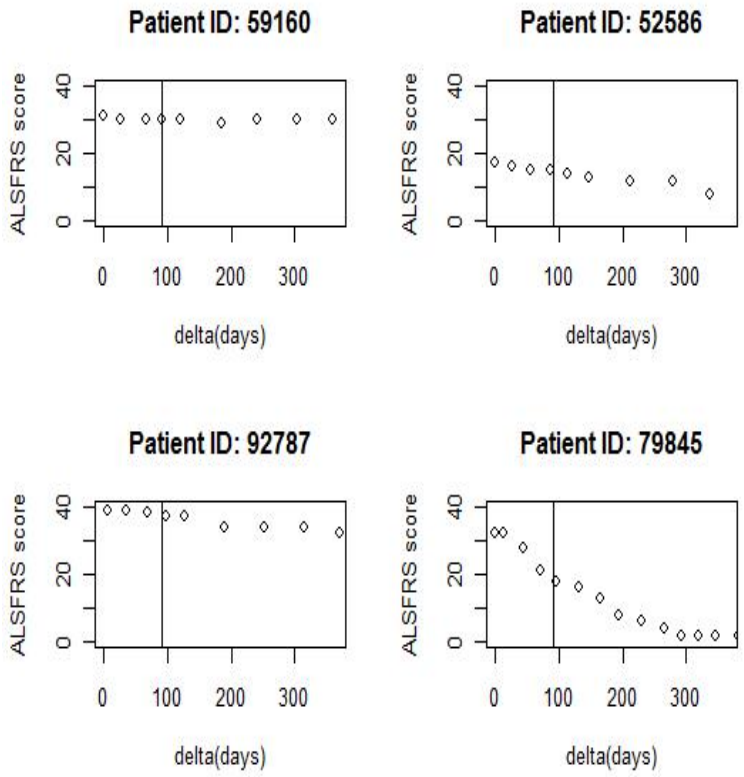
Variable	Found By
onset location	Gordon, 2010 and Qureshi, 2006
age of onset	Gordon, 2010
time between onset of symptoms and diagnosis	Qureshi, 2006
uric acid level	Pagnoni 2012
functional vital capacity (FVC) slope	Magnus, 2002, Traynor, 2004, Kollewe, 2008 and Pagnoni, 2012
body mass index (BMI) and absolute weight	Pagnoni, 2011
variation of heart rate	Pinto 2012

2) Exploratory analysis like plotting and simple linear regression helped identify several interesting predictors/relationship.

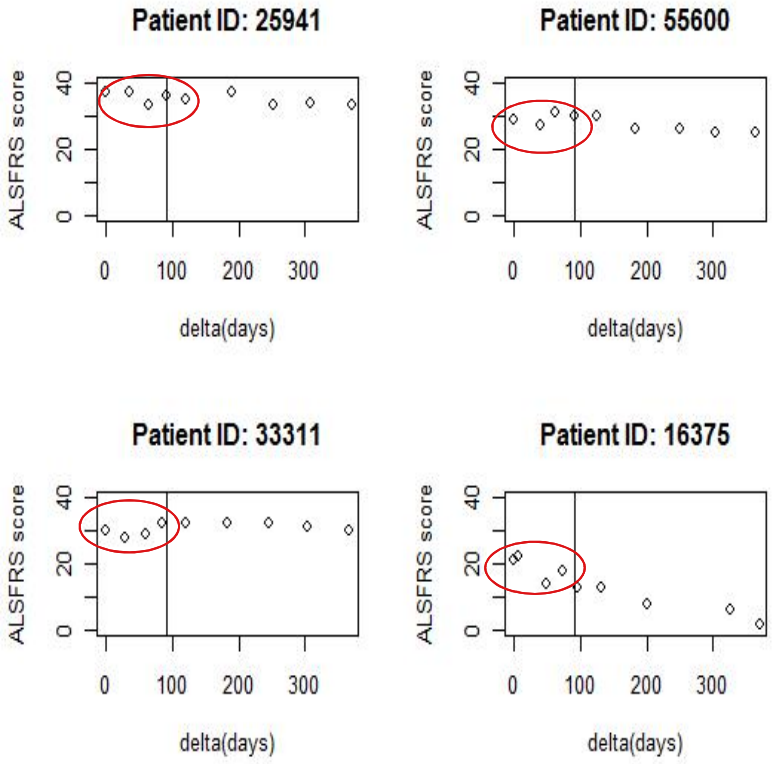
Data Preparation: Exploratory Analysis

2. The ALSFRS history in the first 3 months is informative in predicting future progression, but can also be noisy sometime.

Consistency Between First Three Month Change and Future Change



Noise of ALSFRS Score in First Three Month



Data Preparation: Exploratory Analysis

SENTRANA

3. The score of 10 questions can be grouped into five parts (face, hand, body, leg and chest) based on their functional muscle areas and score correlation.

Score Correlation of Each 10 Questions in ALSFRS

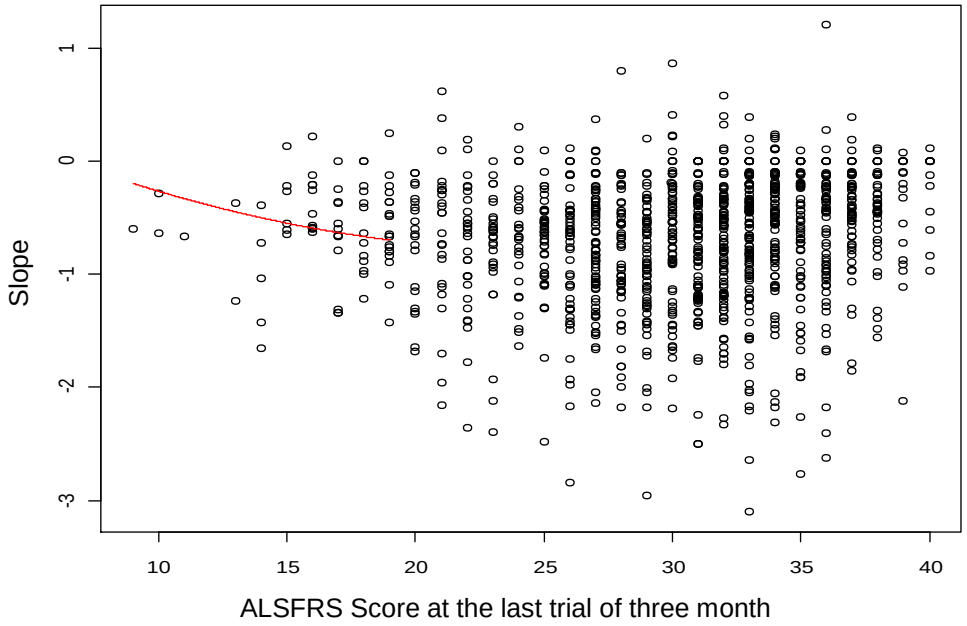
		Face			Hand		Body		Leg		Chest
		Q1: Speech	Q2: Savation	Q3: Swallowing	Q4: Handwriting	Q5: Cutting	Q6: Dressing Hygiene	Q7: Turning In Bed	Q8: Walking	Q9: Climbing_Stairs	Q10: Respiratory
Face	Q1: Speech	1.0000	0.7701	0.7663	0.1373	0.1556	0.0729	0.1104	-0.0167	-0.0034	0.3577
	Q2: Savation	0.7701	1.0000	0.6889	0.0793	0.0995	0.0150	0.0456	-0.0596	-0.0541	0.3251
	Q3: Swallowing	0.7663	0.6889	1.0000	0.1865	0.2312	0.1775	0.1939	0.0801	0.0696	0.4401
Hand	Q4: Handwriting	0.1373	0.0793	0.1865	1.0000	0.8383	0.7332	0.6260	0.2620	0.3085	0.2890
	Q5: Cutting	0.1556	0.0995	0.2312	0.8383	1.0000	0.8106	0.6898	0.3102	0.3519	0.3156
Body	Q6: Dressing Hygiene	0.0729	0.0150	0.1775	0.7332	0.8106	1.0000	0.7897	0.5062	0.5253	0.2938
	Q7: Turning In Bed	0.1104	0.0456	0.1939	0.6260	0.6898	0.7897	1.0000	0.6271	0.6160	0.3266
Leg	Q8: Walking	-0.0167	-0.0596	0.0801	0.2620	0.3102	0.5062	0.6271	1.0000	0.8366	0.2365
	Q9: Climbing_Stairs	-0.0034	-0.0541	0.0696	0.3085	0.3519	0.5253	0.6160	0.8366	1.0000	0.2019
Chest	Q10: Respiratory	0.3577	0.3251	0.4401	0.2890	0.3156	0.2938	0.3266	0.2365	0.2019	1.0000

Data Preparation: Exploratory Analysis

SENTRANA

4. Nonlinear relationship was found between initial disease condition (ALSFRS score of the last trail in the first three months) and the future progression slope.

Quadratic model between ALSFRS score at last trial of three month with the slope



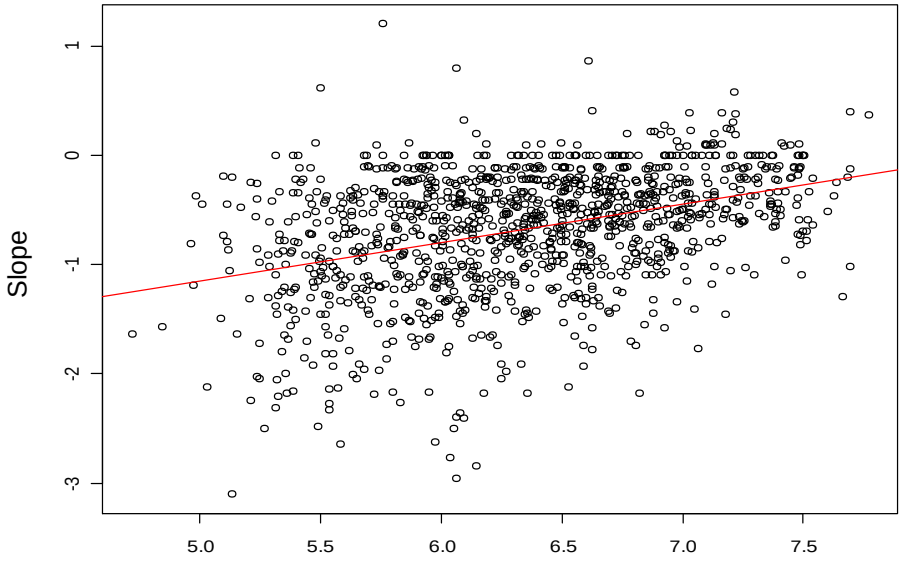
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.6802	0.0161	-42.3647	0.0000
poly(score_three_month, 2)1	1.9057	0.5555	3.4305	0.0006
poly(score_three_month, 2)2	3.1264	0.5555	5.6278	0.0000

Data Preparation: Exploratory Analysis

5. **Time from disease onset is a strong predictor for future progression rate.**

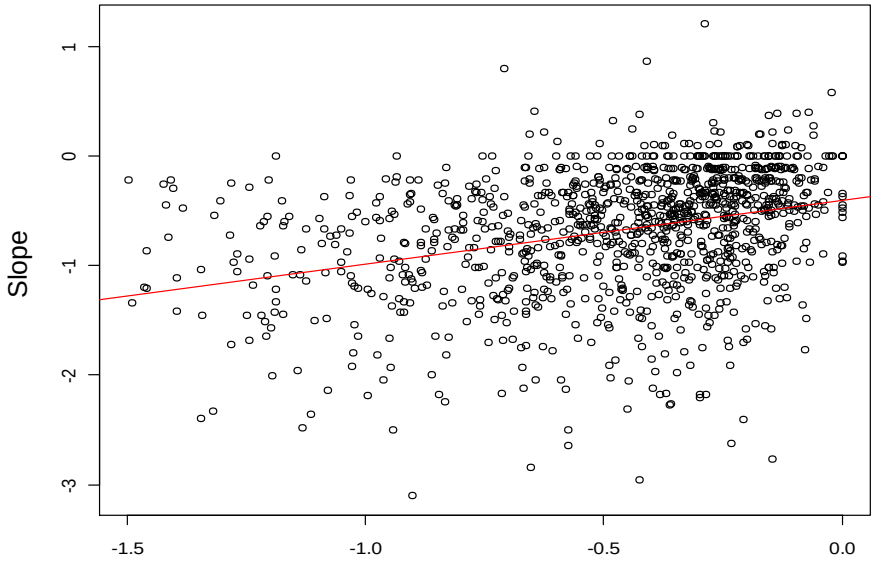
6. **The deterioration rate from onset till now is also very significant for the prediction of future progression rate.**
   (40-ALSFRS Score at First Trial)/ Date Difference Between Onset and First Trail

Plot of slope with log(onset days)



Log Value of time (days) from disease onset

Plot of slope with deterioration rate from onset



(40-ALSFRS Score at First Trial)/ Date Difference Between Onset and First Trail

Data Preparation: Final Variable Definition List

1. **The final variable list used in the next step modeling contains the ALSFRS score information in the first three month...**

Variable Category	Variable Name	Definition Logic
Three Month ALSFRS Score Information	Three_m_slope	Data exploration
	MSE_three_m_slope	Data exploration
	first_score	Data exploration
	score_three_month	Data exploration
	score_three_month_26_abs	Data exploration
	Face	Data exploration
	hand	Data exploration
	Body	Data exploration
	Leg	Data exploration
	Chest	Data exploration
	Face_three_m_slope	Data exploration
	Hand_three_m_slope	Data exploration
	Body_three_m_slope	Data exploration
	Leg_three_m_slope	Data exploration
	Chest_three_m_slope	Data exploration

Data Preparation: Final Variable Definition List (cont')

2. **… and also the information about onset, family ALS history, vital capacity, lab test and conditions for vital signs.**

Variable Category	Variable Name	Definition Logic
Onset Information	Onset_age	Literature and Data exploration
	Onset_delta_log	Literature and Data exploration
	Onset_diag_delta_diff	Literature and Data exploration
	Diag_delta_missing_ind	Missing Value
	Onset_site	Literature
	Onset_slope	Data exploration
Family History	Als_ind	Literature
	Neuro_disease_ind:	Literature
Vital Capacity Information	fvc_slope	Literature and Data exploration
	svc_slope	Literature and Data exploration
	vc_slope:	Inconsistent measures
	fvc_ind	Inconsistent measures
Lab Test	Uric_acid_value:	Literature and Data exploration
	Uric_acid_value_missing_ind	Missing Value
Vital Signs	Weight_slope:	Data exploration
	RR_slope:	Data exploration
	Avg_RR:	Data exploration
	Pulse_slope:	Literature and Data exploration
	Average_pulse:	Literature and Data exploration

3. **Also define dummy variables to take care of variables with large portion of missing values and inconsistent measures.**

Model Specification: Random Forest

1. Introduction to Random Forest Algorithm

1) Random forest is an ensemble classifier that consists of many independent decision trees and the predicted class that is the unweighted average of the values predicted by each of individual trees.
2) Each tree is grown as follows:
   a) If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
   b) If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
   c) Each tree is grown to the largest extent possible. There is no pruning.

2. Advantage of Random Forest

1) One of the most accurate learning algorithms available.
2) Effective to identify and handle the interaction among different predictor variables.
3) Straightforward measurement of variable importance based on the deduction of certain measure of errors for each variable (eg: mean square error).

3. Disadvantage of Random Forest and Solutions

1) Relative low graphic interpretability
   Solutions: Combined with exploratory analysis
2) Potential over fit problems for some datasets with noisy classification/regression tasks
   Solution: Cross validation

Model Specification: Mixed Model

1. Rationales to Use Mixed Models Rather Than A Single Model

1) Important predictors have large percentages of missing values

Percentage of Missing Values for Uric Acid Test Value and Diagnosis Delta Data

Variables	Missing Data Percentage
Uric acid test value	87.22%
Time difference between onset and diagnosis	77.86%

2) Inconsistent measurement.

Coefficients Difference Between FVC Change rate and SVC Change Rate on Slope by Running Separate Linear Models

	Estimate	Std. Error	t value	Pr(>|t|)
FVC Change				
(Intercept)	-0.619162085	0.017965907	-34.46316831	3.60E-176
fvc_slope	0.677254327	0.101618355	6.664685004	4.22E-11
SVC Change				
(Intercept)	-0.348064179	0.078039143	-4.460123047	1.94E-05
svc_slope	4.515342901	0.590007253	7.653029475	7.14E-12

Model Specification: Mixed Model

2. Specification for Mixed Models

1) Use cross validation for variable selection to avoid over fitting.

Mixed Random Forest Specification

Model	Model Description	Variable Used After Variable Selection	Data Used and number of obs
Based Model	Train the model and use the binary indicator to indicate whether certain variable is missing or not	three_m_slope, score_three_month, first_score , score_three_month_26_abs, face, bod, hand, leg, chest, face_three_m_slope, hand_three_m_slope, body_three_m_slope, leg_three_m_slope, chest_three_m_slope, onset_delta_log, onset_age, onset_site, onset_slope, onset_diag_missing_ind, als_ind, neuro_disease_ind, fvc_slope, weight_slope, weight_percentage_slope, uric_acid_value, uric_acid_value_missing_ind	Full data, 1197 obs.
Sub Model 1	Trained by the data when fvc data is available	three_m_slope, score_three_month, first_score, score_three_month_26_abs,face, body, hand, leg, chest,face_three_m_slope, hand_three_m_slope, body_three_m_slope, leg_three_m_slope, chest_three_m_slope, onset_delta_log, onset_age, onset_site, onset_slope,onset_diag_missing_ind,als_ind, neuro_disease_ind,fvc_slope, weight_slope, weight_percentage_slope,uric_acid_value, uric_acid_value_missing_ind	Subset of data when fvc information is available 1082 obs
Sub Model 2	Trained by the data when uric_acid_value is available	three_m_slope, score_three_month_26_abs, onset_delta_log ,onset_diag_delta_diff, onset_diag_missing_ind, onset_site, vc_slope, uric_acid_value	Subset of data when uric_acid_value is available, 153 obs
Sub Model 3	Trained by the data when diagnosis delta data is available	three_m_slope, score_three_month_26_abs, face, body, hand, leg, chest, face_three_m_slope, hand_three_m_slope, body_three_m_slope, leg_three_m_slope, chest_three_m_slope, onset_delta_log, onset_age, onset_diag_delta_diff, vc_slope, avg_pulse + uric_acid_value, uric_acid_value_missing_ind	Subset of data when diagnosis delta data is available, 265 obs

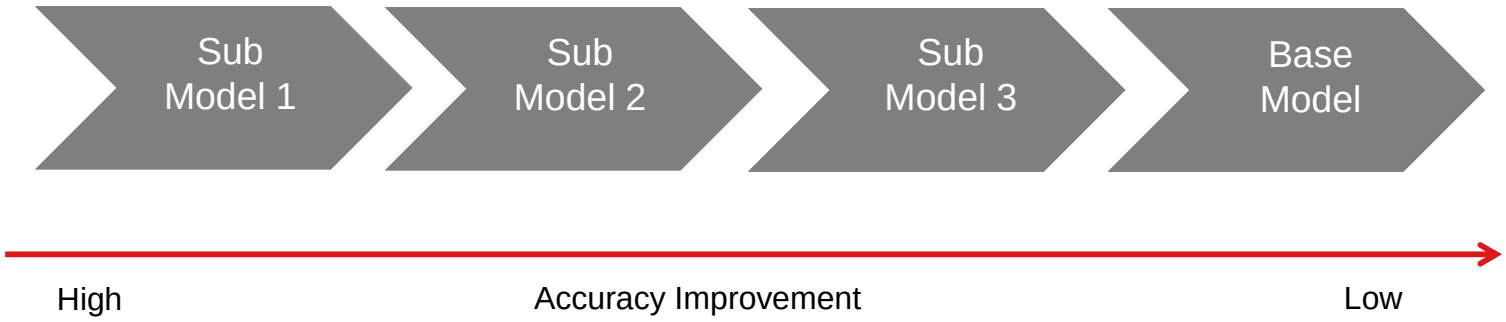
Model Specification: Mixed Model

3. Improvement of Sub Models Over Base Models

RMSD Reduction of Each Sub Model Compared with Base Model

Sub Model	Subset of Test Data	Reduce of RMSD	Number of Obs in the Subset of the validation data
Sub Model1	when fvc information is available	0.010	253
Sub Model2	when uric_acid_value is available	0.005	33
Sub Model3	when diagnosis delta data is available	0.003	59

4. Combine each model



Analysis and Results

1. **Accuracy.** RMSD of 0.512 compared to the RMSD of 0.53 of the baseline algorithm.

2. **No Over Fitting.** Consistent accuracy level both on the test data and on the blinded validation data.

3. More new variables and their relationships with progression rate are found.

4. **Variable Importance**

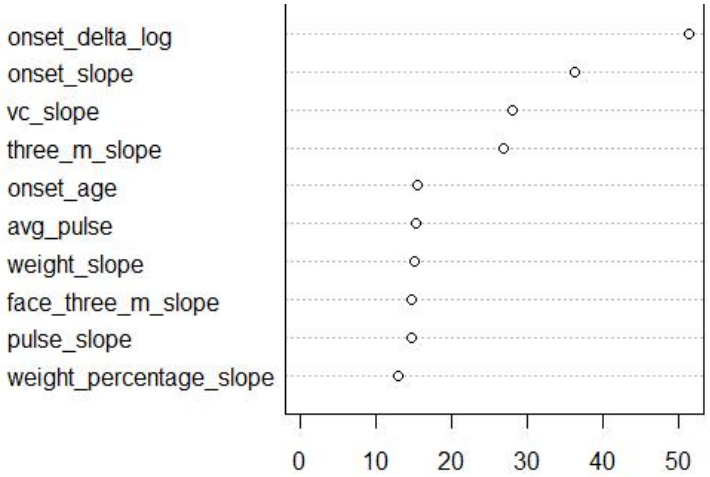
1) **Most important variables:**
   - **length of time for onset**
   - **the deterioration rate from onset till now**
   - **the vital capacity change rate**
   - **the historical ALSFRS score change rate**
   - **the pulse change rate, the weight change rate**
   - **the initial score of ALSFRS**
   - **the historical deterioration rate of face score**

2) Importance of variable such as uric acid value and time difference between onset and diagnosis increase in sub model compared with base model.

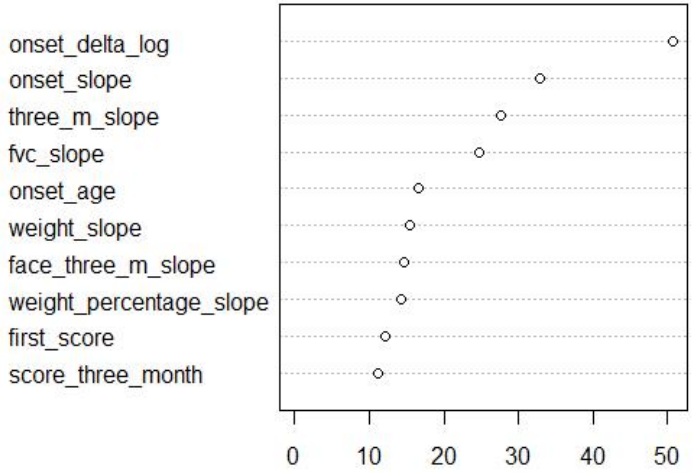
SENTRANA

3. Variable Importance (cont')

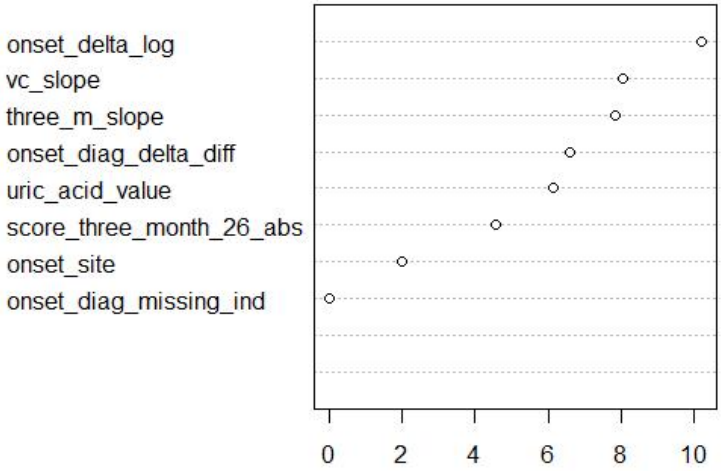
Variable Importance for Base Model



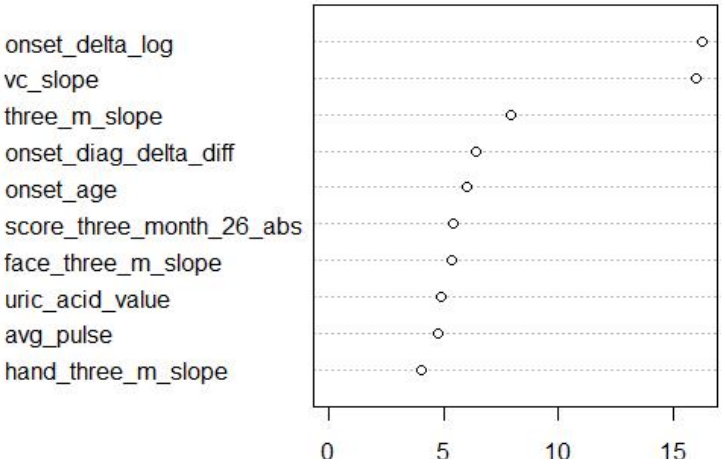
Variable Importance for Sub Model 1 (when fvc information is available)



Variable Importance for Sub Model 2 (when uric acid value is available)



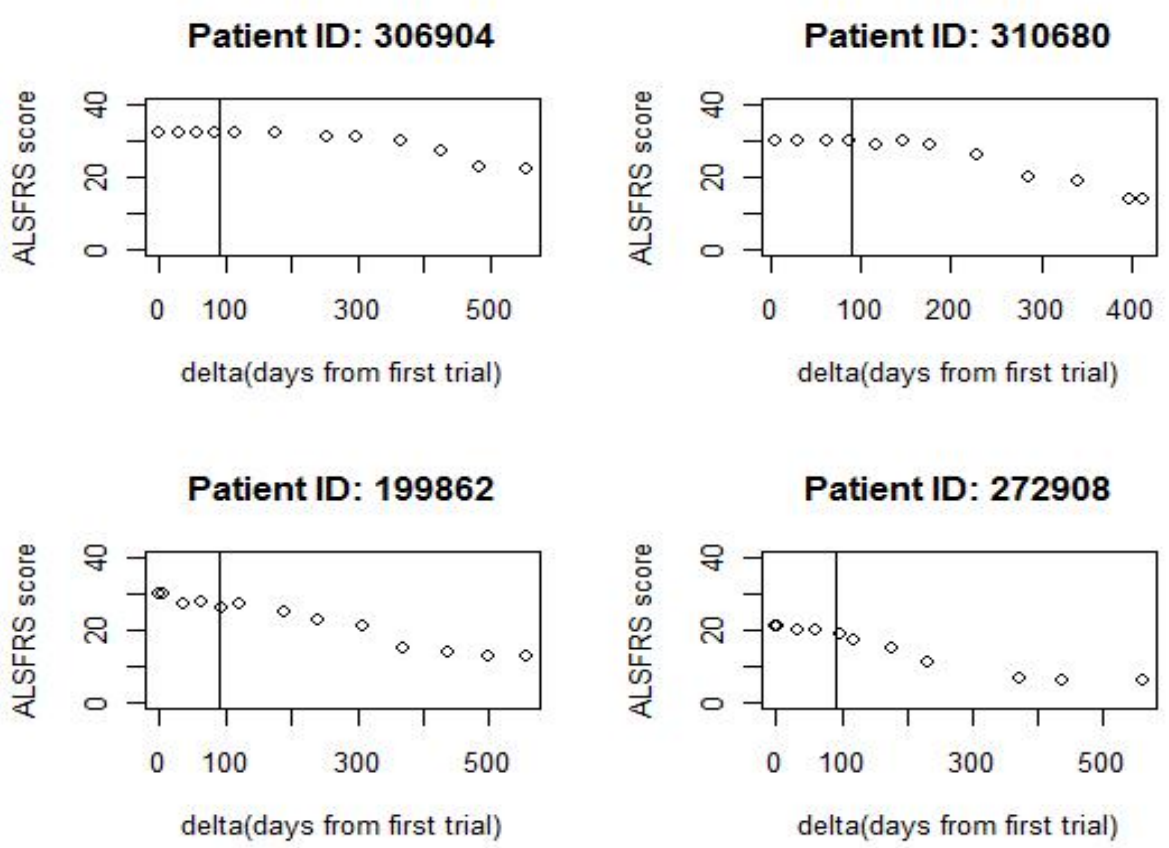
Variable Importance for Sub Model 3 (when diagnosis delta is available)



Interesting Implications

1. **Different stages in ALS progression? Is there heterogeneity for the length of stages for different patients?**

Some Patients with progression stages



Interesting Implications

SENTRANA

2. Look at progression for different body parts rather than combined? Reason for the correlation between body part? Transmission mechanism?

Score Correlation of Each 10 Questions in ALSFRS

		Face			Hand		Body		Leg		Chest
		Q1: Speech	Q2: Savation	Q3: Swallowing	Q4: Handwriting	Q5: Cutting	Q6: Dressing Hygiene	Q7: Turning In Bed	Q8: Walking	Q9: Climbing_Stairs	Q10: Respiratory
Face	Q1: Speech	1.0000	0.7701	0.7663	0.1373	0.1556	0.0729	0.1104	-0.0167	-0.0034	0.3577
	Q2: Savation	0.7701	1.0000	0.6889	0.0793	0.0995	0.0150	0.0456	-0.0596	-0.0541	0.3251
	Q3: Swallowing	0.7663	0.6889	1.0000	0.1865	0.2312	0.1775	0.1939	0.0801	0.0696	0.4401
Hand	Q4: Handwriting	0.1373	0.0793	0.1865	1.0000	0.8383	0.7332	0.6260	0.2620	0.3085	0.2890
	Q5: Cutting	0.1556	0.0995	0.2312	0.8383	1.0000	0.8106	0.6898	0.3102	0.3519	0.3156
Body	Q6: Dressing Hygiene	0.0729	0.0150	0.1775	0.7332	0.8106	1.0000	0.7897	0.5062	0.5253	0.2938
	Q7: Turning In Bed	0.1104	0.0456	0.1939	0.6260	0.6898	0.7897	1.0000	0.6271	0.6160	0.3266
Leg	Q8: Walking	-0.0167	-0.0596	0.0801	0.2620	0.3102	0.5062	0.6271	1.0000	0.8366	0.2365
	Q9: Climbing_Stairs	-0.0034	-0.0541	0.0696	0.3085	0.3519	0.5253	0.6160	0.8366	1.0000	0.2019
Chest	Q10: Respiratory	0.3577	0.3251	0.4401	0.2890	0.3156	0.2938	0.3266	0.2365	0.2019	1.0000

Thank you !
Q&A

Nov 13, 2012

Appendix: Detailed variable definition

Variable Category	Variable Name	Description
Three Month ALSFRS Score Information	Three_m_slope	The slope is defined as the coefficients of a fitted line of the ALSFRS score with ALSFRS_Delta in the first three month.
	MSE_three_m_slope	The Measn Standard Error of the fitted line of the first three month score.
	first_score	The total sum of ALSFRS_score for the first trial.
	score_three_month	The total sum of ALSFRS_score for the las trial where delta<91
	score_three_month_26_abs	=abs(score_three_month-26). Score_three_month is found to be nonlinear correlated with slope. When score_three_month>26, it is positive correlated and when score_three_month<26, it is negatively correlated.
	Face	The score of the questions that is related to the face part of the musle (Speech_1,Salivation_2,Swallowing_3) in the last trial during the first month
	hand	The score of the questions that is related to the hand part of the musle (Handwriting_4, Cutting_wo_Gastrostomy_5a or Cutting_wo_Gastrostomy_5b) in the last trial during the first month
	Body	The score of the questions that is related to the body part of the musle (Dressing_Hygiene_6,Turning_in_Bed_7) in the last trial during the first month
	Leg	The score of the questions that is related to the leg part of the musle (Walking_8,Climbing_Stairs_9) in the last trial during the first month
	Chest	The score of the questions that is related to the chest part of the musle (Respiratory or R_3_Respiratory_Insufficiency) in the last trial during the first month
	Face_three_m_slope	The slope of face scores in the first three month, where slope is defined as: slope=(score for first trial-score for last trial)/(delta_first_trial-delta_first_trial)
	Hand_three_m_slope	The slope of hand scores in the first three month, where slope is defined as: slope=(score for first trial-score for last trial)/(delta_first_trial-delta_first_trial)
	Body_three_m_slope	The slope of body scores in the first three month, where slope is defined as: slope=(score for first trial-score for last trial)/(delta_first_trial-delta_first_trial)
	Leg_three_m_slope	The slope of leg scores in the first three month, where slope is defined as: slope=(score for first trial-score for last trial)/(delta_first_trial-delta_first_trial)
	Chest_three_m_slope	The slope of chest scores in the first three month, where slope is defined as: slope=(score for first trial-score for last trial)/(delta_first_trial-delta_first_trial)

Appendix: Detailed variable definition (cont')

Variable Category	Variable Name	Description
Onset Information	Onset_age	The age at onset, it is defined as onset_age=age+onset_delta/365
	Onset_delta_log	=log(-onset_delta)
	Onset_diag_delta_diff	The difference between diagnosis delta with onset delta
	Onset_site	Categorical Varaible with three possible values "Bulbar" , "Limb", "Both"
	Onset_slope	onset_slope=(40-first_score)/onset_delta*365.24/30
Family History	Als_ind	If there is any family member with ALS history before
	Neuro_disease_ind:	If there is any family member with any neurogology disease before.
Vital Capacity Information	fvc_slope	The slope is defined as the coefficients of a fitted line of the fvc value with fvc delta in the first three month.
	svc_slope	The slope is defined as the coefficients of a fitted line of the svc value with svc delta in the first three month.
	vc_slope:	vc slope=fvc slope if fvc data is availabe and svc_slope otherwise
	fvc_ind	binary indicator to indicate if the vc_slope variable is from fvc or svc. fvc=ind when vc slope is from fvc, otherwise 0.
Lab Test	Uric_acid_value:	the average value of uric_acid during the first three month
	Uric_acid_value_missing_ind:	If the uric_acid_value is missing for this patients
Vital Signs	Weight_slope:	The slope is defined as the coefficients of a fitted line of the weight value with date delta in the first three month.
	RR_slope:	The slope is defined as the coefficients of a fitted line of the respiratory rate with the date delta in the first three month.
	Avg_RR:	Average respiratory rate in the first three month.
	Pulse_slope:	The slope is defined as the coefficients of a fitted line of the pulse rate with the date delta in the first three month.
	Average_pulse:	Average pulse rate value in the first three month.

Appendix: Significance and Effect of Selected Variables

Significance of Selected Variables and Effect on Response Variables By Using a Simple Linear Regression

Predictors	Effect	P-Value
ALSFRS slope of first three month	Steeper change rate in first three month, quicker progression rate	1.38e-13 ***
Delta (days) of between onset and first trial	Larger time difference, slower progression rate	<2e-16 ***
The score change from onset till the first trial	Larger change rate, quicker progression rate	<2e-16 ***
ALSFRS score of last trial in first three month	Nonliear	6.01e-11 ***
Uric Acid	Larger uric acid values, slower progression rate	0.0106 *
FVC Change	Steeper FVC changes, quicker progression rate	4.22e-11 ***
SVC Change	Steeper SVC changes quicker progression rate	1.66e-13 ***
The score of last trial related to chest area muscle	Smaller score for the question related to muscle in chest area, quicker progression rate	0.00136 **
The score of last trial related to face area muscle	Smaller score for the question related to muscle in facet area, quicker progression rate	4.41e-13 ***
Time Difference Between Onset and Diagnosis	Larger the time difference, the slower progression rate	0.105

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1