Model-based and Model-free Techniques for Amyotrophic Lateral Sclerosis Diagnostic Prediction and Patient Clustering

Ming Tang 1,2 \* , Chao Gao 1,2 \*, Stephen A. Goutman 3 \*, Alexandr Kalinin 1,4, Bhramar Mukherjee 2, Yuanfang Guan 4, Ivo D. Dinov 1,4,5

1 Statistics Online Computational Resource,

Department of Health Behavior and Biological Sciences,

2 Department of Biostatistics,

3 Department of Neurology,

4 Department of Computational Medicine and Bioinformatics,

5 Michigan Institute for Data Science,

University of Michigan, Ann Arbor, MI 48109

**Supplementary Materials**

**Methods**

**Prediction Models**

**Multiple Linear Regression:** As an extension of simple linear regression, a multiple linear regression model consists of at least two explanatory variables and one response variable. The relationship between predictors and the outcome variable is modeled by fitting a linear equation. If we assume that the covariates are orthonormal, i.e., $X^{T}X=I$, then the ordinary least squares (OLS) estimates minimize

$$\min\_{β\in R^{p}}\left\{\frac{1}{N}\left∥y-Xβ\right∥\_{2}^{2}\right\}.$$

Then, the OLS solution is defined by$ ^^{OLS}=(X^{T}X)^{-1}X^{T}y=X^{T}y.$ We used step-wise variable selection in our linear modeling (using the R *stat* package and the function $step()$). Specifically, the step-wise model relies on the Akaike Information Criterion (AIC) to apply synchronous “backward” and “forward” feature selection. If the covariates are not orthonormal, regularization methods may employed to obtain a solution based on jointly minimizing the fidelity and the regularization terms of the cost function, see below.

**Random Forest:** Random forests [1] is an ensemble learning method for classification and regression by growing a number of random decision trees and averaging the outputs, with a goal to correct for the overfitting problem. R package *Caret* [2] and 5-fold cross validation is employed for training process to build optimal model. The Random Forest Algorithm iteratively (for $b = 1, 2, 3, …, B$) draws a bootstrap sample $Z^{\*}$ of size $N$ from the training data. For each terminal node of the tree, the algorithm recursively grows a random-forest tree $T\_{b}$ to the bootstrapped data until the minimum node size $n\_{min}$ is reached. At each step, we select $m$ features at random from the $k$ variables, pick the best variable/split-point among the $m$ features, and split the node into two children nodes. The output includes the ensemble of trees $\{T\_{b}\}\_{b=1}^{B}$. The random forest prediction at a new point $x$ may be obtained by regression:

1. $\hat{f}\_{rf}^{B}\left(x\right)=\frac{1}{B}\sum\_{b=1}^{B}T\_{b}\left(x\right).$

If $C\_{b}^{}(x)$ is the class prediction of the $b$th random forest classifier, then $\hat{C}\_{rf}^{B}\left(x\right)$ is the majority vote $\left\{\hat{C}\_{b}^{}\left(x\right)\right\}\_{b=1}^{B}$.

**Bayesian Additive Regression Tree (BART):** BART [3] is a nonparametric approach that uses dimensionally adaptive random basis element. It tackles the sum-of-trees model by applying a prior that regularizes the fit by keeping the individual tree effects small. R package *BARTMachine* [4] is used to calculate optimal training model through 5-fold cross-validation. BART is an ensemble method where the response variable $Y$ is estimated by a sum of Bayesian Classification and Regression Trees (CART). For an $n×k$ design matrix of predictors, $X=[x\_{1}$,$ x\_{2},x\_{3},…, x\_{k}]^{T}$, where $x\_{i}$ =($x\_{i,1},x\_{i,2},x\_{i,3},…x\_{i,k},$) represents the *ith* row (*ith* observation), the BART model is:

$$Y\_{i}=\sum\_{j=1}^{m}g(x\_{i} |T\_{j},M\_{j})+ϵ\_{i},$$

where $g(x\_{i})$ is a Bayesian CART decision tree model, $T\_{j}$ is a decision tree with a terminal node parameter $M\_{j}$, $j=1,2,3,…,m$, $m$ is the total number of trees in the model, $ϵ\_{i}, \~ N(0, σ^{2})$, and $σ^{2}$ is the residual variance. To compare the alternative prediction methods, we compared the results of the corresponding ALSFRS change predictions using several validation measures including coefficient of determination (R2), root-mean-square error (RMSE), and Pearson’s correlation coefficient ($ρ$), representing the linear bivariate association between the *predicted* and *actual* ALSFRS change.

Except for linear regression modeling, the model-free machine learning predictions are not derived from ordinary least squares. Therefore, the standard coefficient of determination is not applicable here. Instead, alternative pseudo R2 measures may be used to assess the proportion of the variance of the slope that is actually predicted.

**Evaluation Metrics**

**Pearson’s correlation coefficient:** For paired observations $X=\{x\_{1},x\_{2},x\_{3},...,x\_{n}\}$ and $Y=\{y\_{1},y\_{2},y\_{3},...,y\_{n}\}$, both containing $n$ values, like observed ($X$) and predicted ($Y$) ALSFRS slope change for all patients, the correlation coefficient is computed by

$$ρ=ρ\_{X,Y}=\frac{\sum\_{i=1}^{n}\left(x\_{i}y\_{i}\right)-n\overbar{x}\overbar{y}}{\sqrt{\left(\sum\_{i=1}^{n}x\_{i}^{2}-n\overbar{x}^{2}\right)} \sqrt{\left(\sum\_{i=1}^{n}y\_{i}^{2}-n\overbar{y}^{2}\right)}}$$

Where $\overbar{x}$ and $\overbar{y}$ represent the sample arithmetic averages for the $X$ and $Y$ vectors.

**Coefficient of determination:** The correlation coefficient is defined by $R^{2}=1-\frac{RSS}{TSS}$, where the *total*, $TSS$ (proportional to the variance in the data) and regression, $RSS$ (explained by a linear model) sums of squares are defined by: $TSS=\sum\_{i=1}^{n}\left(y\_{i}-\overbar{y}\right)^{2}$ and $RSS=\sum\_{i=1}^{n}\left(ε\_{i}\right)^{2}=\sum\_{i=1}^{n}\left(y\_{i}-\hat{y\_{i}}\right)^{2}$, where $\hat{y\_{i}} $ is the estimate of the predictive value, and$y\_{i} $is the true value of the predicted subject. Various adjusted $R^{2}$ metrics have been proposed to account automatic and spurious increase of $R^{2}$ when the number of explanatory model variables increases. And the **Root Mean Square Error** is defined by

$RMSE=\sqrt{\frac{1}{n}\sum\_{i=1}^{n}(y\_{i}-\hat{y\_{i}})^{2}}$.

**Feature Selection Strategy**

Two methods, random forest and knockoff filter, are applied for selecting important predictors for the outcome of interest, e.g., ALSFRS\_slope.

**Random forests (RF):** For the purpose of feature selection, RF carries out the following algorithm. (1) Compute mean-square-error ($MSE\_{0}$) for current model (2) For $j$th feature in the model: permute feature $j$; compute $MSE\_{j}$ for the new model; (3) Calculate percentage increase $MSE$ (%IncMSE). A large %IncMSE implies the feature is important. Then the features are ranked accordingly, and top 20 of them are selected.

**Knockoff filter (KO):** It is a novel method for feature selection in high dimensional data with false discovery rate (FDR) control. Generally, the procedure involves three steps (1) Construct knockoffs (null features) of original covariates, with the same covariance matrix; (2) Calculate appropriate statistics for original and knockoff feature pairs. Lasso is well known for its capability of feature selection with accuracy. Hence, the statistics are based on lasso model; (3) Set a threshold for the statistics obtained in the last step with a target FDR. Given the statistics and threshold, the features can be selected.

For each imputed dataset (20 in total), there are X iterations for RF and Y iterations for KO. The number of times for each features being selected are tracked, and further transformed into the proportions. In this way, the results from RF and KO are comparable. The final selected features are those stand out in both methods.

**Multiple Imputation**

Other than list-wise deletion, mean/median substitution or single imputation, multiple imputation fills in data properly without changing any relationships within data. $m-$chain multiple imputation imputes $m$ values for each missing cell and create $m$ “complete” data matrixes. This multiple imputation enables the estimates of the missing cells to be efficient and not biased too far from the true value. By doing it $m$ times, we are able to reflect the uncertainty of the estimates of the missing cell.

A prior studies suggests that the bootstrapping-based Amelia II algorithm speeds up the computation and generates similar results compared to imputation via EM algorithm. The better computational performance enables Amelia to deal with larger dataset with ease. Because of the strength of computation, in this study, we used the “Amelia” R packages ([https://cran.rproject.org/web/vpackages/Amelia/vignettes/ amelia.pdf](https://cran.rproject.org/web/vpackages/Amelia/vignettes/%20amelia.pdf)) to conduct the multiple imputation to our high-dimension data.

**Data Source and Pre-processing**

Data management: The original PRO-ACT training dataset contains clinical and lab test information of 8,635 patients. The progression of ALS is measured by the slope of the Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) score. We used information of only 2,424 subjects that had valid ALSFRS slopes, which measure the symptoms severity and patients’ functional status. ALSFRS is a checklist of ten different assessments of motors skills, of which the degeneration is the most predominant symptom of ALS. Every element of the checklist measures the ability ranging from 0 to 4 (4 corresponds to normal function). Therefore, the ALSFRS total score ranges from 0 to 40.

The time points for all longitudinally varying data elements were aggregated into signature vectors of *(min, median, max, slope*) minimum, median, maximum values and the slope, for each data element, respectively. A new feature “*ALSFRS\_slope\_init*” is computed to represent the ALSFRS slope change over the first three months, which will be used to fit or train the models and then validate the performance on the 12-month follow up.

Aggregate of longitudinal data: Aside from the ALSFRS slope, some changing demographic features, and several other disease characteristics (e.g., *Race, gender, age, onset\_delta, onset\_site, if\_use\_riluzole*), out of 49 features, 42 of them were longitudinally recorded for most subjects at different time points, see **Supplementary** **Table S.1**. We refer to these longitudinal covariates as “features with change”. We tried to integrate as much of the temporal information as possible using several alternative methods. However, due to the heterogeneous time points of observation, we opted to synthesize a moderate number of covariates (4) representing features with change.

For each patient and each longitudinal feature, a four-tuple signature vector was derived by calculating the *minimum, median, maximum* and *the slope* of the specific data element. The slope was fitted by a linear model of the feature value and the time of observation:

$y\_{feature\\_value}=a+β\_{slope}×x\_{time}$.

For example, **Supplementary Table S.1** shows nine observations of *systolic blood pressure (bp\_systolic)* for patient SubjectID: 533, extracted directly from the raw PRO-ACT dataset. Then the derived signature vector would be $\left(min, median,max, slope\right) ≡ \left(129, 139, 160, -0.01673\right)$, where the slope of the change of the feature value is computed by:

$\hat{β}\_{slope}=\frac{\sum\_{i=1}^{5}(x\_{i}- \overbar{x})(y\_{i}-\overbar{y})}{\sum\_{i=1}^{5}\left(x\_{i}- \overbar{x}\right)^{2}}=-0.01673$.

After synthesizing the longitudinal data elements into four-tuple signature vectors, we were able to harmonize and aggregate the training and testing datasets. The training dataset included Subject\_IDs and 171 features for a total of 2,424 patients, and the testing data had 101 subjects with the same 171 features, paired with SubjectID.

**Table S.1**Error! No text of specified style in document.**:** Example of longitudinal feature synthesis.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Time (day) | 0 | 12 | 25 | 39 | 53 | 91 | 189 | 287 | 378 |
| Bp\_systolic | **142** | **139** | **129** | **160** | **140** | **138** | **134** | **141** | **134** |

**Results**

**Supplementary** **Figure S.1**. Raw data missing pattern prior to imputation.



**Supplementary Table S.2**

**Table S.2**: Temporally static variables vs. longitudinally varying features.

| Index | x | Type | Label |
| --- | --- | --- | --- |
| 1 | Gender | Constant | Categorical |
| 2 | Age | Constant | Continuous |
| 3 | Race | Constant | Categorical |
| 4 | onset\_delta | Constant | Continuous |
| 5 | onset\_site | Constant | Categorical |
| 6 | if\_use\_Riluzole | Constant | Categorical |
| 7 | ALSFRS\_Total | longitudinal |  |
| 8 | mouth | longitudinal |  |
| 9 | hands | longitudinal |  |
| 10 | trunk | longitudinal |  |
| 11 | leg | longitudinal |  |
| 12 | respiratory | longitudinal |  |
| 13 | bp\_diastolic | longitudinal |  |
| 14 | bp\_systolic | longitudinal |  |
| 15 | pulse | longitudinal |  |
| 16 | BMI | longitudinal |  |
| 17 | fvc\_percent | longitudinal |  |
| 18 | Monocytes | longitudinal |  |
| 19 | Chloride | longitudinal |  |
| 20 | AST(SGOT) | longitudinal |  |
| 21 | CK | longitudinal |  |
| 22 | White Blood Cell (WBC) | longitudinal |  |
| 23 | Glucose | longitudinal |  |
| 24 | Alkaline Phosphatase | longitudinal |  |
| 25 | Basophils | longitudinal |  |
| 26 | Calcium | longitudinal |  |
| 27 | Hemoglobin | longitudinal |  |
| 28 | Platelets | longitudinal |  |
| 29 | Sodium | longitudinal |  |
| 30 | Blood Urea Nitrogen (BUN) | longitudinal |  |
| 31 | Potassium | longitudinal |  |
| 32 | Lymphocytes | longitudinal |  |
| 33 | Red Blood Cells (RBC) | longitudinal |  |
| 34 | Protein | longitudinal |  |
| 35 | Phosphorus | longitudinal |  |
| 36 | ALT(SGPT) | longitudinal |  |
| 37 | Albumin | longitudinal |  |
| 38 | Hematocrit | longitudinal |  |
| 39 | Bicarbonate | longitudinal |  |
| 40 | Creatinine | longitudinal |  |
| 41 | Eosinophils | longitudinal |  |
| 42 | Neutrophils | longitudinal |  |
| 43 | Urine Ph | longitudinal |  |
| 44 | Gamma-glutamyltransferase | longitudinal |  |
| 45 | HbA1c (Glycated Hemoglobin) | longitudinal |  |
| 46 | Total Cholesterol | longitudinal |  |
| 47 | Triglycerides | longitudinal |  |
| 48 | Bilirubin (total) | longitudinal |  |

**Supplementary** **Table S.3**.

**Table S.3**: Descriptive statistics of an instance of the complete dataset following multivariate imputation.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Name | Mean | Standard Deviation | Median |
| fvc\_percent3 | 78.49114 | 21.59163 | 79.1614 |
| fvc\_percent2 | 79.8769 | 20.23648 | 80.31362 |
| fvc2 | 3.213604 | 1.12762 | 3.14 |
| Q5a\_Cutting\_without\_Gastrostomy | 2.740861 | 1.209612 | 3 |
| Potassium | 4.197128 | 0.663593 | 4.2 |
| Chloride | 103.5307 | 3.222934 | 104 |
| R1\_Dyspnea | 3.692861 | 0.641383 | 4 |
| Q1\_Speech | 3.269596 | 0.994067 | 4 |
| Red Blood Cells (RBC) | 49623018 | 4.78E+08 | 4700 |
| Q4\_Handwriting | 2.94756 | 1.104178 | 3 |
| bp\_diastolic | 81.19322 | 10.41474 | 80 |
| Q9\_Climbing\_Stairs | 2.072415 | 1.429784 | 2 |
| White Blood Cell (WBC) | 6.880742 | 2.273702 | 6.6 |
| Q8\_Walking | 2.743375 | 0.969817 | 3 |
| trunk | 5.697867 | 1.911129 | 6 |
| Protein | 72.43594 | 4.498393 | 72 |
| Monocytes | 6.460268 | 2.52749 | 6.2 |
| temperature | 36.56529 | 1.479417 | 36.6 |
| ALSFRS\_R\_Total | 38.33876 | 5.355327 | 39 |
| fvc3 | 3.160067 | 1.167622 | 3.1 |
| Urine Ph | 5.704764 | 0.795907 | 6 |
| Creatinine | 69.86755 | 18.01122 | 70.72 |
| R3\_Respiratory\_Insufficiency | 3.97173 | 0.239178 | 4 |
| diag\_delta | -285.651 | 272.0575 | -203 |
| pulse | 75.76844 | 11.13198 | 76 |
| Q2\_Salivation | 3.444718 | 0.8689 | 4 |
| Platelets | 236.5106 | 56.79195 | 231 |
| Neutrophils | 64.04545 | 9.034688 | 64.3 |
| fvc\_percent | 81.95934 | 17.88929 | 82.54941 |
| fvc | 3.431412 | 1.080327 | 3.36 |
| respiratory\_rate | 17.46235 | 3.400895 | 17 |
| Glucose | 5.570063 | 1.692252 | 5.3 |
| bp\_systolic | 131.3522 | 16.53222 | 130 |
| weight | 75.90628 | 15.83524 | 75 |
| hands | 5.684879 | 2.187968 | 6 |
| Lymphocytes | 26.40233 | 7.531528 | 26 |
| Q6\_Dressing\_and\_Hygiene | 2.617571 | 1.100967 | 3 |
| Q10\_Respiratory | 3.702866 | 0.514372 | 4 |
| Q7\_Turning\_in\_Bed | 3.079933 | 0.959833 | 3 |
| Q5\_Cutting | 2.737728 | 1.212448 | 3 |
| fvc1 | 3.4402 | 1.075976 | 3.37 |
| Calcium | 2.340433 | 0.16047 | 2.3453 |
| Sodium | 139.7302 | 2.611484 | 140 |
| R2\_Orthopnea | 3.873023 | 0.393713 | 4 |
| Lactate Dehydrogenase | 170.3398 | 48.91598 | 164 |
| ALSFRS\_Total | 30.19189 | 5.725022 | 31 |
| Hematocrit | 41.88932 | 7.364655 | 43 |
| Eosinophils | 2.513298 | 1.757029 | 2.1 |
| Q3\_Swallowing | 3.569756 | 0.692 | 4 |
| respiratory | 3.714409 | 0.520805 | 4 |
| leg | 4.815789 | 2.295667 | 4 |
| mouth | 10.28393 | 2.281076 | 11 |
| fvc\_normal | 4.181074 | 0.937172 | 4.201414 |
| Albumin | 44.05963 | 3.261636 | 44 |
| Partial Thromboplastin Time | 27.05032 | 3.097204 | 26.9 |
| CK | 314.4097 | 305.5316 | 223 |
| Hemoglobin | 143.9247 | 12.98412 | 144 |
| Basophils | 0.508305 | 0.443953 | 0.4 |
| Urine Protein | 42.45455 | 28.46529 | 30 |
| Bicarbonate | 27.0562 | 3.045853 | 27 |
| respiratory\_R | 11.53761 | 0.960085 | 12 |
| ALT(SGPT) | 35.05641 | 27.05596 | 30 |
| fvc\_percent1 | 82.18476 | 17.75817 | 82.7883 |
| Bilirubin (Total) | 10.62317 | 7.293688 | 10 |
| Phosphorus | 1.184171 | 0.172945 | 1.1951 |
| AST(SGOT) | 30.67548 | 20.69509 | 28 |
| Alkaline Phosphatase | 78.81415 | 29.18413 | 74 |
| Prothrombin Time (clotting) | 28.5841 | 33.22694 | 12 |
| Urine Glucose | 12.73333 | 83.80609 | 0 |
| Blood Urea Nitrogen (BUN) | 5.628877 | 1.854202 | 5.4765 |
| Bilirubin (Direct) | 1.024222 | 1.371856 | 0 |
| Absolute Basophil Count | 0.159462 | 0.766126 | 0.03 |
| Absolute Eosinophil Count | 0.214121 | 0.358301 | 0.14 |
| Triglycerides | 2.068257 | 1.514356 | 1.65 |
| Absolute Lymphocyte Count | 1.786619 | 0.587581 | 1.7 |
| Total Cholesterol | 5.891553 | 1.107439 | 5.83 |
| Absolute Monocyte Count | 0.448453 | 0.177428 | 0.43 |
| HbA1c (Glycated Hemoglobin) | 5.416156 | 0.727117 | 5.4 |
| BMI | 0.002592 | 0.000447 | 0.002533 |
| Absolute Neutrophil Count | 4.478812 | 1.523459 | 4.23 |
| Urine blood | 21.75 | 13.20038 | 25 |
| Gamma-glutamyltransferase | 36.97673 | 48.18865 | 26 |
| height | 170.7452 | 9.686642 | 171 |
| Urine Albumin | 0.034314 | 0.20773 | 0 |
| Urine WBCs | 5.256757 | 7.815247 | 2 |
| Beta HCG | 2.15 | 2.521449 | 1 |
| Band Neutrophils | 0.596512 | 1.165734 | 0 |
| Urine Ketones | 13.46154 | 12.64658 | 5 |
| svc\_percent | 85.08871 | 14.2141 | 84.39268 |
| svc\_normal | 4.207625 | 0.922328 | 4.187889 |
| svc | 3.594384 | 1.033539 | 3.48 |
| Segmented Neutrophils | 68.0459 | 10.08571 | 70 |
| Absolute Band Neutrophil Count | 4.146957 | 3.906902 | 4.95 |
| Absolute Segmented Neutrophil Count | 5.233333 | 1.886716 | 5 |
| International Normalized Ratio (clotting) | 1.026486 | 0.093513 | 1.01 |
| Urine RBCs | 4.075 | 5.227143 | 2.5 |
| Thyroid Stimulating Hormone | 2.012908 | 1.240649 | 1.8 |
| Uric Acid | 292.1184 | 84.9518 | 297.4 |
| Free T3 | 0.004591 | 0.000943 | 0.0045 |
| Urine Specific Gravity | 1.010151 | 0.011076 | 1 |
| Amylase | 66.87179 | 26.38179 | 63 |
| Free T4 | 13.60156 | 4.560657 | 14.2 |
| Q5b\_Cutting\_with\_Gastrostomy | 2.473404 | 1.411106 | 3 |
| IMMUNOGLOBULIN A | 233.2231 | 83.96033 | 227.5 |
| Fibrinogen | 332.124 | 62.4197 | 327 |
| GAMMA-GLOBULIN | 1.053846 | 0.258545 | 1 |
| ALPHA1-GLOBULIN | 0 | 0 | 0 |
| BETA-GLOBULIN | 1.099237 | 1.135815 | 1 |
| Mean Corpuscular Hemoglobin | 30.4099 | 1.414641 | 30 |
| Mean Corpuscular Hemoglobin Concentration | 36.87047 | 31.16219 | 34 |
| IMMUNOGLOBULIN G | 1.053846 | 0.226587 | 1 |
| C-Reactive Protein | 1.36 | 2.494944 | 1 |
| Mean Corpuscular Volume | 90.38083 | 3.907751 | 90 |
| Albumin/globulin ratio | 1.4 | 0.491793 | 1 |
| IMMUNOGLOBULIN M | 147.4462 | 109.3181 | 113 |
| ALPHA2-GLOBULIN | 1 | 0 | 1 |
| Bilirubin (Indirect) | 32.4976 | 4.525293 | 34.208 |
| Urine Hyaline Cast | 2.5 | 0.707107 | 2.5 |
| Total T4 | 10.4 | 0.424264 | 10.4 |
| T3 Uptake | 101 | 7.071068 | 101 |
| Creatine Kinase MB | 91 | 100.9356 | 55 |

**References**

1. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

2. Kuhn, M., *Caret package.* Journal of Statistical Software, 2008. **28**(5): p. 1-26.

3. Chipman, H.A., E.I. George, and R.E. McCulloch, *BART: BAYESIAN ADDITIVE REGRESSION TREES.* Annals of Applied Statistics, 2010. **4**(1): p. 266-298.

4. Kapelner, A. and J. Bleich, *bartMachine: Machine Learning with Bayesian Additive Regression Trees.* Journal of Statistical Software, 2016. **70**(4): p. 1-40.