Neuroinformatics (2019) 17:407–421
https://doi.org/10.1007/s12021-018-9406-9

ORIGINAL ARTICLE

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]

Published online: 20 November 2018

Abstract

Amyotrophic lateral sclerosis (ALS) is a complex progressive neurodegenerative disorder with an estimated prevalence of about 5 per 100,000 people in the United States. In this study, the ALS disease progression is measured by the change of Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) score over time. The study aims to provide clinical decision support for timely forecasting of the ALS trajectory as well as accurate and reproducible computable phenotypic clustering of participants. Patient data are extracted from DREAM-Phil Bowen ALS Prediction Prize4Life Challenge data, most of which are from the Pooled Resource Open-Access ALS Clinical Trials Database (PRO-ACT) archive. We employed model-based and model-free machine-learning methods to predict the change of the ALSFRS score over time. Using training and testing data we quantified and compared the performance of different techniques. We also used unsupervised machine learning methods to cluster the patients into separate computable phenotypes and interpret the derived subcohorts. Direct prediction of univariate clinical outcomes based on model-based (linear models) or model-free (machine learning based techniques – random forest and Bayesian adaptive regression trees) was only moderately successful. The correlation coefficients between clinically observed changes in ALSFRS scores relative to the model-based/model-free predicted counterparts were 0.427 (random forest) and 0.545(BART). The reliability of these results were assessed using internal statistical cross validation and well as external data validation. Unsupervised clustering generated very reliable and consistent partitions of the patient cohort into four computable phenotypic subgroups. These clusters were explicated by identifying specific salient clinical features included in the PRO-ACT archive that discriminate between the derived subcohorts. There are differences between alternative analytical methods in forecasting specific clinical phenotypes. Although predicting univariate clinical outcomes may be challenging, our results suggest that modern data science strategies are useful in clustering patients and generating evidence-based ALS hypotheses about complex interactions of multivariate factors. Predicting *univariate* clinical outcomes using the PRO-ACT data yields only marginal accuracy (about 70%). However, unsupervised clustering of participants into sub-groups generates stable, reliable and consistent (exceeding 95%) computable phenotypes whose explication requires interpretation of multivariate sets of features.

Highlights

- Used a large ALS data archive of 8,000 patients consisting of 3 million records, including 200 clinical features tracked over 12 months.
- Employed model-based and model-free methods to predict ALSFRS changes over time, cluster patients into cohorts, and derive computable phenotypes.

Ming Tang, Chao Gao and Stephen A. Goutman contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12021-018-9406-9) contains supplementary material, which is available to authorized users.

✉ Ivo D. Dinov
statistics@umich.edu; https://www.umich.edu/~dinov

Extended author information available on the last page of the article

- Research findings include stable, reliable, and consistent (95%) patient stratification into computable phenotypes. However, clinical explication of the results requires interpretation of multivariate information.

Keywords ALS · Amyotrophic lateral sclerosis · Decision support · Machine learning · Predictive analytics · Data science · Big data

Introduction

Prior clinical studies pooling heterogeneous datasets across sites and sources have been successful in examining, simulating and modeling disease progression patterns, e.g., studies of Alzheimer's disease (Saykin et al. 2015; Moon et al. 2015a, b) and Parkinson's disease (Dinov et al. 2016; Marek et al. 2011). Recently, similar effort has been made to integrate all available Amyotrophic Lateral Sclerosis (ALS) clinical information. The Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) (https://nctu.partners.org/ProACT/) (Zach et al. 2015) database collects and aggregates clinical data of 16 ALS clinical trials and one observational study completed in the recent twenty years (Atassi et al. 2014). These big data initiatives to better characterize ALS are ongoing and have the potential to provide insight into disease progression and heterogeneity. While PRO-ACT includes clinical variables on a large number of participants, other programs such as Answer ALS are incorporating "omics" and even wearables data. As ALS enters the era of big data, we utilized the PRO-ACT dataset to help explore the differences in analytical techniques, potential pitfalls, and how these methods can be leveraged in the future. These initiatives, especially when pulled from electronic medical record systems, require acquiring raw data, computational preprocessing, aggregation, harmonization, analysis, and visualization leading to clinically relevant result interpretation. We report on building such workflow that is tested to examine complex ALS data. Next, we employ model-based methods and model-free machine learning techniques for automated disease progression tracking in ALS patients and discuss the relative strengths/weakness of each.

Statistical Model-Based vs. Data Science Model-Free Methods

Model-based and model-free techniques represent complementary approaches for predictive big data analytics (Dinov 2016). The former class of techniques includes classical parametric models under some of the following specific assumptions: time series analyses assume model parameters are constant over time, residuals are homoscedastic, and stationarity over some order, and use historic or training data to capture trends or patterns and extrapolate prospective or forecast future observations. Other examples of model-based methods include (1) multivariate regression methods (Chatterjee and Hadi 2015), which represent variable interdependencies between predictors and responses in terms of some base functions (e.g., polynomials) whose coefficients capture the influence of all variables on the outcomes and facilitate forward predictions, (2) generalized estimating equations models (Bergsma et al. 2009) is a semi-parametric technique for population-average inference, rather than subject-specific inference in the linear models, and structural equation models (Markus 2012), which assume asymmetric causal relationships among normally distributed interval-level measured variables. In general, model-based methods are contingent on many (parametric) assumptions that may not be satisfied in many situations, particularly with heterogeneous sources of data.

Alternatively, model-free machine-learning techniques (Edwards et al. 2009), classification theory (Zhang 2000), network analysis (Kai-Hsiang et al. 1999), and hierarchical clustering (Filzmoser et al. 1999) may be employed for unsupervised data mining (Witten and Frank 2005), hyperparameter tuning (Wistuba et al. 2015; Saitta et al. 2010), pattern recognition (De Sa 2012), trend identification (Tamás Kincses et al. 2008), k-means (Jain 2010) or fuzzy (Wismüller et al. 2004) clustering. Contrary to the reductionist approach of reality-simplification employed in the model-based methods, model-free techniques combine, evolve, ensemble and train algorithms that adapt to the contextual affinities of a process learning the dynamic characteristics that may explain the observations. Just like with cognitive and social brain development, maturation and aging, model-free techniques may need to be constantly adjusted and are not guaranteed to always be correct. However, when properly trained, effectively maintained, and continuously reinforced, model-free learning based methods provide powerful and reliable forecasting for complex phenomena.

Model-based statistical estimation relies on training data to estimate concrete parameters, like model coefficients, that are explicitly assumed to be included in a specific analytical framework representing a proxy of the underlying process (e.g., GEE, SEM, LASSO, etc.) In contract, model-free inference does not rely on a predetermined analytical framework as a fixed stereotypic representation of the problem. Rather, model-free techniques adapt their structure to the characteristics of the specific data they emulate. For instance, in this context, random forests and decision trees are model-free techniques, as the topology of the graphical networks, as well as the actual branching rules, are freely determined by the data affinities, instead of being prescribed in advance by a specific graph model. Similarly, whereas some clustering techniques are model-based, e.g., Gaussian mixture models, others do not have an *apriori* structure and rely simply on data affinities, manifold distance measures, or entropy information criteria to group cases into clusters.

In principle, no technique, algorithm, or an approach is really completely model-free, as there are always underlying axiomatic assumptions or characterizations that are expected in all scientific settings. However, the degree to which a specific technique relies on a strict rendition of its analytical representation does divide model-based analyses from model-free inferential approaches. Solving constrained or unconstrained linear models is rather different from solving non-convex machine-learning optimization problems (Gong et al. 2013; Bubeck 2015; Mairal 2015). The solutions to many model-based problems are largely consistent and reproducible, and can be obtained in polynomial time (Fiedler 2006). Model-free solutions to non-convex problems may be NP-hard, frequently rely on heuristics, and may be unstable (Jain and Kar 2017; Allen-Zhu and Hazan 2016).

Prior ALS Predictive Modeling Studies

In a previous DREAM-Phil Bowen ALS Prediction Prize4Life crowdsourcing challenge (Kuffner et al. 2015), teams of investigators developed several promising predictive models based on the clinical database PRO-ACT (Zach et al. 2015). The modeling methods introduced by these teams include Bayesian trees, Random Forest, and Nonparametric regression, which outperformed support vector regression (baseline method), multivariate regression, and linear regression. Notably, among teams implementing the same machine learning algorithm, the accuracy and reliability of the predictive models was influenced by the ways they preprocessed the longitudinal data in the raw dataset. Unlike the time-static features, all time points for longitudinal clinical measurements may be transformed by linear regression into a signature vector of congruent measurements, e.g., four-tuple like correlation, variability, slope and intercept of a temporal linear model.

Random Forest (Rodriguez-Galiano et al. 2012) is popular modeling method. In this PROACT dataset, the decline in ALSFRS score is highly correlated with changes in other time series, and there could be interaction among different body parts while the ALS is worsening. Random Forest is considered to be good at dealing with predictor variables of high co-linearity and interaction. However, the graphic interpretability is relatively low. Moreover, overfitting does exist and it can be improved by cross-validation (Carreiro et al. 2015; Grigull et al. 2016).

Some prior studies used the approximated classification expectation-maximization (CEM) algorithm to obtain clusters of patients based on 6 features (*onset_delta, fvc_percent, ALSFRS_Total, weight, ALSFRS speech* (*Q1*), and *Trunk muscle*), see https://www.synapse.org/#!Synapse:syn4957429/wiki/237065. Another group employed CART-based clustering techniques (Steinberg and Colla 2009) to split the patients into 7 clusters using various clinical outcomes, including *ALSFRS_slope*. This classification approach generates results that are more intuitive to interpret compared to classical methods such as PCA, see https://www.synapse.org/#!Synapse:syn4942470/wiki/235991.

The previous investigations using PRO-ACT data to predict ALS progression or forecast survival have demonstrated mostly marginal prediction accuracy. For instance, the top-ranked method for predicting survival of individual patients yielded a concordance index of 0.717 (Huang et al. 2017), the optimal RUS-Boost model predicting ALSFRS-R decline generated a cross-validation AUC of 0.82, and the best prediction model for slow progressing patients had an overall prediction accuracy of 0.74 (Ong et al. 2017). Another recent European study examined longitudinally about 2000 patients over two decades. A multivariable Royston-Parmar model was used to predict the composite survival outcome in individual patients (Westeneng et al. 2018). The authors evaluated the model reliability (prediction accuracy 0.77–0.80) using external populations as well as statistically via cross-validation.

This study addresses the following four specific problems:

- Using baseline and 3-month follow-up data, fit models and learn multi-feature affinities to predict the 12-month ALSFRS slope change.
- Identify the most salient PRO-ACT data features that are associated with specific clinical phenotypes.
- Forecast the forced vital capacity (FVC) percent change between 3 and 12 months.
- Use machine learning techniques to derive computed phenotypes clustering the ALS patients.

The main finding of this investigation confirms that predictions of specific univariate clinical ALS outcomes, using the PRO-ACT data, are not expected to be highly accurate, reliable, or consistent. However, unsupervised clustering of the ALS patients into subgroups generates exceptionally stable, reliable and consistent computed phenotypes whose explication as clinically relevant traits requires interpretation of multivariate sets of features.

Methods

We used the PRO-ACT dataset downloaded on June 16, 2017. For each time-resolved feature, four statistics are included. e.g., the maximum and minimum measurement value, delta values for the first and last measurement value, and the slope of the time series. The outcome of interest was the ALS Functional Rating Scale (ALSFRS), original version (Cedarbaum and Stambler 1997), instead of the revised version, ALSFRS-R (Cedarbaum et al. 1999). This choice was motivated by the fact that the majority of PRO-ACT records contained only original ALSFRS scores. We successfully preprocessed and incorporated all categorical variables as binary features or dummy variables, see **Supplementary Table S.2**.

The **Supplementary Materials Methods** section describes the model-based (e.g., ordinary and regularized linear models) and model-free (e.g., Random Forest (RF), Knockoff (KO) filtering, and Bayesian Additive Regression Tree (BART)) techniques. Similarly, the **Data Source and Preprocessing** section in the **Supplementary Materials** includes all details about the data management and manipulations, e.g., the conversion of the longitudinal time varying data elements into signature vectors *(min, median, max, slope*) and missing data imputation. Figure 1 illustrates the end-to-end flowchart of our approach starting with the pre-processed training and testing PRO-ACT data. The protocol includes aggregation of the longitudinal data elements into a signature vector of four components, resolving incomplete data elements, and application of alternative model-based and model-free machine learning prediction techniques. The final external validation of the training-data based models is accomplished using the forecasting results on the separate testing cases. To compare the results of different prediction methods we used the coefficients of determination and correlation.

We employ several machine learning methods for predicting 3–12 month ALSFRS slope of change, without estimating ALSFRS at different time points. By partitioning the open Pooled Resource Open-Access ALS Clinical Trials Database (PRO-ACT) data (www.ALSdatabase.org) (Filzmoser et al. 1999) into training (estimation) and testing (validation) sets, we fit different models to forecast the progression of the disease. We report both positive and negative findings as well as the accuracy and efficacy of the successful approaches to predict clinical phenotypes. In support of open-science principles and to facilitate community-based validation and collaborative revision and expansion of these methods, we are releasing the entire end-to-end protocol, software tools, and research findings.

The **Supplementary Materials** section includes all details about the techniques for data processing, classification and predictive analytics, including random forest clustering and imputation, Bayesian Adaptive Regression Tree (BART) classification, knockoff filtering and feature selection, and appropriate forecasting assessment strategies. Briefly, we generated summary statistics for all raw data elements, selected highly-observed features, imputed the data, computed linear model-based predictions and model-free forecasts of specific clinical outcomes, used unsupervised machine learning methods to obtain derived computable phenotypes, i.e., cluster all patients into subgroups, and finally evaluated all clustering and classification results.

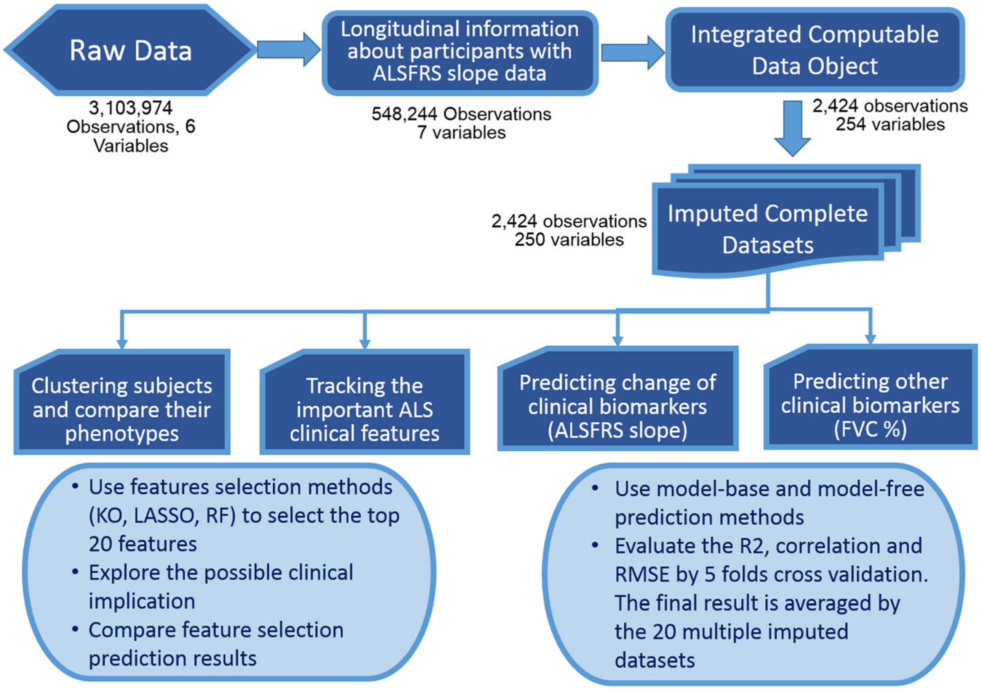

Raw Data
3,103,974 Observations, 6 Variables
Longitudinal information about participants with ALSFRS slope data
548,244 Observations 7 variables
Integrated Computable Data Object
2,424 observations 254 variables
Imputed Complete Datasets
2,424 observations 250 variables
Clustering subjects and compare their phenotypes
Tracking the important ALS clinical features
Predicting change of clinical biomarkers (ALSFRS slope)
Predicting other clinical biomarkers (FVC %)
• Use features selection methods (KO, LASSO, RF) to select the top 20 features
• Explore the possible clinical implication
• Compare feature selection prediction results
• Use model-base and model-free prediction methods
• Evaluate the R2, correlation and RMSE by 5 folds cross validation. The final result is averaged by the 20 multiple imputed datasets

Fig. 1 Diagrammatic depiction of the protocol from raw data preprocessing to the clustering, prediction modeling, results comparison and validation

Data Availability In the spirit of "open science", we have documented, packaged and shared our complete protocol, software tools and scripts used for the data management, processing, aggregation, harmonization, analysis and visualization. These resources (https://github.com/SOCR/ALS_PA) can be used, expanded, tested and validated by the entire community on ALS, alternative neurodegenerative disorders, or other biomedical or health challenges involving large, heterogeneous, incongruent and multi-source datasets.

Results

In this section, we report cohort-based descriptive statistics and selection of salient features, as well as results from model-based predictions of specific clinical outcomes, model-free classification, and computable-phenotype clustering.

Descriptive Statistics A summary of the patient demographics is shown in Table 1**.**

The missing pattern in the data, prior to imputation, is shown on **Supplementary Fig. S.1**. Although it shows substantial missingness in a number of features, the pattern does not suggest strong non-random missingness. After excluding the highly correlated covariates "fvc_normal_min", "fvc_normal_median" and "fvc_normal_slope", 20 imputed datasets were created using 20-chain multiple imputation (Su et al. 2011; Buuren and Groothuis-Oudshoorn 2011; Abayomi et al. 2008). The multiple imputed datasets were homogeneous between different chains. **Supplementary Table S.3** shows an example of the descriptive statistics of a randomly selected complete dataset, the 11th element of the imputation chain.

Table 2 Metrics assessing the machine learning based prediction of ALSFRS slope change, using the coefficient of determination (R^2), root mean square error (RMSE), and the paired correlations (observed vs. predicted ALSFRS slope values)

	R^2	RMSE	Correlation
Random forest	0.1916	0.5633	0.4460
BART	0.2187	0.5537	0.4718

Prediction of 12-Month ALSFRS Slope Change Table 2 shows the results of training RF and BART classifiers on baseline and 3-month follow up data and predicting ALSFRS slope change at 12-month follow up. Random forest and BART methods generated the best supervised machine learning predictions of the ALSFRS slope change. Albeit the correlations between the observed and predicted slope change was around 0.45–0.47, which suggests that this univariate outcome (ALSFRS Total slope change between 3 and 12 months) can't be reliably predicted using the available baseline and 3-month follow up data.

Feature Selection Table 3 depicts the results of the feature selection using knockoff (KO) filtering, left, and random forest (RF), right. Commonly identified features are highlighted. For both feature selection approaches, highly salient features as ranked as influential and placed in the top rows, according of the rank-order of the corresponding raw frequencies and relative proportions. As expected, ALSFRS_Total_slope (baseline to 3 month follow up), onset_delta.x, fvc_percent_slope, and Absolute.Monocyte Count_slope are selected as salient predictors of the 12-month ASLFRS Total slope change. To make the results of RF and KO feature selection comparable and to identify their agreement, we report the relative proportion (Prop) of times each feature is chosen by either algorithm, instead of relying on default measures of variable importance.

Table 1 Basic cohort descriptive statistics

Gender	N	Observed Rate	F	M				
	2424	100%	876(36.14%)	1548(63.86%)				
Onset-site	N	Observed Rate	Bulbar	Limb and Bulbar	Limb			
	2424	100%	494(20.38%)	18(0.74%)	1912(78.88%)			
Race	N	Observed Rate	Asian	Black	Hispanic	Other	Unknown	White
	2424	100%	21 (0.87%)	31 (1.28%)	10 (0.41%)	9 (0.37%)	46 (1.90%)	2307 (95.17%)
Treatment group	N	Observed Rate	Active	Placebo				
	1622	67%	1268(78.18%%)	354(21.82%)				
ALS family history	N	Observed Rate	No	Yes				
	375	15%	296(78.93%)	79(21.07%)				
Riluzole use	N	Observed Rate	No	Yes				
	2060	85%	709(34.42%)	1351(65.58%)				

Table 3 Top 20 features selected by knockoff filtering (left) and random forest (right)

KO			RF		
Features	Freq	Prop	Features	Freq	Prop
treatment_group	364	0.60181	*ALSFRS_Total_slope*	100	1
Age.x	300.5	0.49683	*onset_delta.x*	100	1
Gender	286	0.47285	*fvc_percent_slope*	84.5	0.845
if_use_Riluzole	226	0.37365	weight_slope	83	0.83
Q1_Speech_min	220	0.36373	fvc_slope	56	0.56
pulse_max	215.5	0.3563	bp_systolic_slope	55	0.55
onset_delta.x	208.5	0.34472	fvc_percent1_slope	54	0.54
fvc_percent_slope	165	0.2728	Phosphorus_median	49	0.49
ALSFRS_Total_slope	156.5	0.25875	Sodium_slope	44	0.44
Phosphorus_max	156	0.25792	Creatinine_slope	43	0.43
onset_site	153	0.25296	*Absolute.Monocyte Count_slope*	43	0.43
Q3_Swallowing_min	148.5	0.24552	mouth_slope	42	0.42
Absolute.Monocyte Count_slope	139	0.22981	bp_diastolic_slope	40.5	0.405
fvc_percent1_min	134	0.22155	Bicarbonate_slope	40	0.4
mouth_min	130	0.21493	Absolute.Lymphocyte Count_slope	40	0.4
Absolute.Neutrophil Count_min	118	0.1951	Alkaline Phosphatase_slope	39	0.39
Protein_max	115	0.19013	Red.Blood.Cells RBC._median	39	0.39
fvc_percent_min	110	0.18187	CK_slope	38	0.38
leg_max	108	0.17856	fvc1_slope	38	0.38
leg_slope	108	0.17856	Total Cholesterol_slope	38	0.38

Italicized features are chosen by both feature-selection methods

Forecasting FVC Percent Change Table 4 shows the results of random forest-based prediction of another clinically relevant outcome variable, FVC percent change. The top and bottom rows in the table report two types of change prediction results – either using the baseline FVC as a predictor or not. Clearly the prediction power increases substantially when we include the baseline FVC%, the correlation between observed and predicted FVC%-change increases from 0.67 (no baseline FVC) to 0.83 (with baseline FVC).

We also used supervised classification to group the patients by specific clinical phenotypes, however, the results were not particularly reliable or precise, i.e., the assigned classification labels were not highly associated with specific univariate clinical outcomes.

Unsupervised Clustering Next, we tried unsupervised classification, e.g., k-means clustering, to stratify the participants into separate cohorts. These results were interesting as clustering all patients into four groups allowed us to examine deeper the four computed phenotypes (machine learning derived phenotypic subgroups). Figure 2 shows a plot of the objective function capturing the within-group error rate across increasing number of clusters. The rapid error rate decrease until the elbow (corresponding to 4 clusters) suggests *four* as a reasonable number of computable phenotypes present in this ALS cohort. The clinical interpretation of these four computable phenotypes requires special attention as the clusters do not necessarily correspond to univariate clinical outcomes. Figure 3 illustrates ALSFRS total score changes over time

Table 4 Metrics assessing the machine learning prediction of FVC change

	R^2	RMSE	Correlation
Using baseline FVC	0.6817	14.27	0.8293
Without Baseline FVC	0.4308	19.10	0.6671

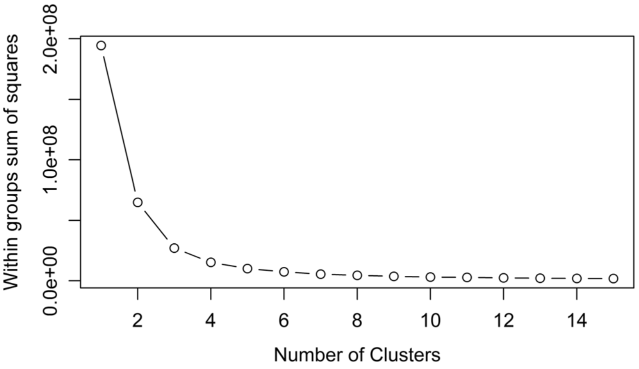


Fig. 2 Unsupervised clustering identified four phenotypic subgroups

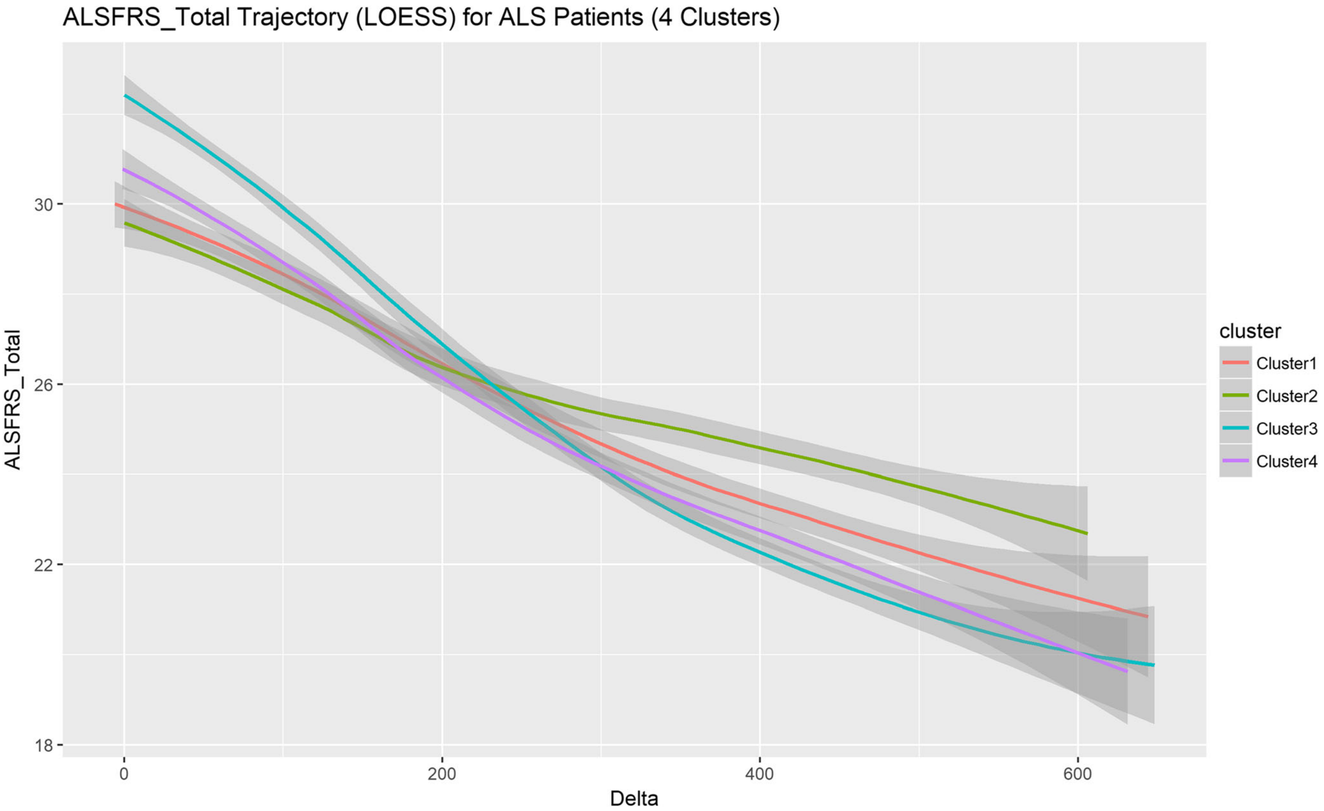


Fig. 3 Plot of the locally weighted scatterplot smoothing (LOESS) models of the overall observed ALSFRS Total score trajectories for the patients in the four phenotypic subgroups. Confidence bands are drawn for each trajectory to illustrate the expected dispersion at each delta – time point (in days). The fairly tight packing of the trajectories for all four clusters suggests that the computed clinical phenotypes are not distinguishable with respect to a single clinical outcome feature, e.g., ALSFRS Total score

for each of these machine learning derived sub-cohorts. Notice the intertwined nature of these trajectories suggesting that the four computed phenotypes blend a mix of participants with a widely varying distribution of this specific clinical outcome, in this case ALSFRS Total score over time (horizontal axis).

As neither the model-based nor the model-free classification results showed strong coherence with the specific univariate clinically relevant outcome (ALSFRS Total score change), we examined the *consistency* and *reliability* of automated unsupervised clustering. Table 5 demonstrates the regularity and dependability of unsupervised clustering into four categories across 1000 randomly-initialized clustering iterations. The reported mean proportions (consistency) and standard deviations (reliability) indicate the robustness of the patients' clustering across all experiments. The large mean proportions (mean~1) and small dispersions (SD~0) suggest that patients remain largely and consistently grouped within the same cohorts, with very few patients transitioning from one group to another across the repeated automated unsupervised re-clustering. This suggests reliable and consistent automated machine learning clustering of ALS patients into four computable phenotypic sub-cohorts. The interpretations of these machine learning derived computed phenotypes (clusters) requires a deeper exploration of multiple factors, rather than examining simple univariate clinical outcomes. The size of each of the four

Table 5 Reliability and consistency of machine-learning based ALS classification

Class ID	Mean Proportions	SD
1	1	0
2	0.986	0.018
3	0.956	0.053
4 (unclassified)	0.985	0.018

Table 6 Cluster sizes and average silhouette width values. Silhouette values measure the similarity of participants within their own cluster (cluster cohesion), relative to other phenotypes (cluster separation), $-1 \leq Silhouette \leq +1$ with low or high values indicating poor or reliable matching of participants within their phenotypic sub-group, respectively

Cluster ID (cluster-size)	Cluster 1 (565)	Cluster 2 (427)	Cluster 3 (699)	Cluster 4 (733)
Average silhouette widths	0.581	0.626	0.503	0.500

Table 7 Main ALS variables associated with significant between-cluster differences

Feature name	Between cluster significant differences					
	1–2	1–3	1–4	2–3	2–4	3–4
onset_site			1	1	1	
onset_delta.x	1	1	1	1	1	1
onset_delta.y	1	1	1	1	1	
Red.Blood.Cells..RBC._min	1			1	1	
Red.Blood.Cells..RBC._median	1			1	1	
Red.Blood.Cells..RBC._slope				1	1	
Q4_Handwriting_max		1		1	1	
Q4_Handwriting_min		1		1	1	
Q4_Handwriting_median		1		1	1	
Q9_Climbing_Stairs_max		1	1	1	1	
Q9_Climbing_Stairs_min		1	1	1	1	
Q9_Climbing_Stairs_median		1	1	1	1	
Q9_Climbing_Stairs_slope	1			1		
Q8_Walking_max		1	1	1	1	
Q8_Walking_min		1	1	1	1	
Q8_Walking_median		1	1	1	1	
trunk_max		1	1	1	1	1
trunk_min		1	1	1	1	
trunk_median		1	1	1	1	
Protein_slope	1			1	1	
Creatinine_max		1	1	1		
Creatinine_min		1	1	1	1	
Creatinine_median		1	1	1	1	
respiratory_rate_max	1			1	1	
hands_max		1		1	1	
hands_min		1		1	1	
hands_median		1		1	1	
Q6_Dressing_and_Hygiene_max		1	1	1	1	
Q6_Dressing_and_Hygiene_min		1		1	1	
Q6_Dressing_and_Hygiene_median		1	1	1	1	
Q7_Turning_in_Bed_max		1	1	1	1	
Q7_Turning_in_Bed_min		1		1	1	
Q7_Turning_in_Bed_median		1	1	1	1	
Sodium_slope	1			1	1	
ALSFRS_Total_max		1		1	1	1
ALSFRS_Total_min		1		1		1
ALSFRS_Total_median		1		1	1	1
ALSFRS_Total_slope				1	1	
Hematocrit_max	1			1	1	
Hematocrit_min	1			1	1	
Hematocrit_median	1			1	1	
leg_max		1	1	1	1	
leg_min		1	1	1	1	
leg_median		1	1	1	1	
mouth_min			1		1	1
Absolute.Basophil.Count_max	1			1	1	
Absolute.Basophil.Count_min	1			1	1	

Table 7 (continued)

Feature name	Between cluster significant differences					
	1–2	1–3	1–4	2–3	2–4	3–4
Absolute.Basophil.Count_median	1			1	1	
Absolute.Basophil.Count_slope	1			1	1	
Absolute.Eosinophil.Count_max	1			1	1	
Absolute.Eosinophil.Count_median	1			1	1	
Absolute.Eosinophil.Count_slope	1			1	1	
Absolute.Lymphocyte.Count_slope	1			1	1	
Absolute.Monocyte.Count_slope	1			1	1	

clusters and their corresponding average silhouette width values are shown on Table 6. Clearly the unsupervised clustering splits the patients in four balanced sub-cohorts representing separate computable phenotypes. The relatively high silhouette values for all phenotypes suggest that the grouping of subjects into clusters is consistent and reliable.

Table 7 shows the main features associated with significant statistical differences between clusters. Cell indicators "1" and "" (blank values) denote respectively the significant and insignificant differences between a pair of clusters (column) for a specific variable (row). Generally speaking, we only include the most wide-spread cluster discriminants, skipping over features, among the 171, that may have discriminated only between one pair of clusters or none at all. Note that the interpretation of clinical and phonotypic differences between a pair of clusters *I* and *J*, e.g., 2–3 (column 5), typically involves a number of features. Explication of the derived subcohorts complicates the objective understanding of the four computable phenotypes. This reflects the reality that ALS is an extremely heterogeneous progressive neurodegenerative disorder with multifaceted clinical manifestations. Selecting a very small number of features is unlikely to yield highly predictive models. Our results hint to a more nuanced computable phenotypic motifs whose elucidation requires joint interpretation of multiple clinical features.

These findings illustrate that clinical interpretations of unsupervised machine learning clustering results, representing derived or computed clinical phenotypes, require joint inspection and holistic review of multiple data features. For example, Table 7 shows that the *trunk_max* feature was an important discriminant between the four computable phenotypes, as it played a vital role in separating all pairs of sub-cohorts, except the first two. Whereas, another feature, *Q9_Climbing_Stairs_slope*, only discriminated between cohorts 1 and 2, as well as 2 and 3, but not among the others. Thus, in this reliable and consistent unsupervised clustering of patients into four computable phenotypic groups, the explication of clinical relevance relies on high-dimensional multivariable interpretation.

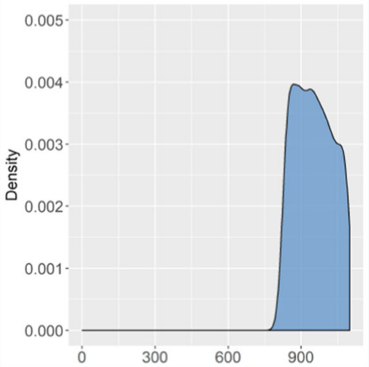
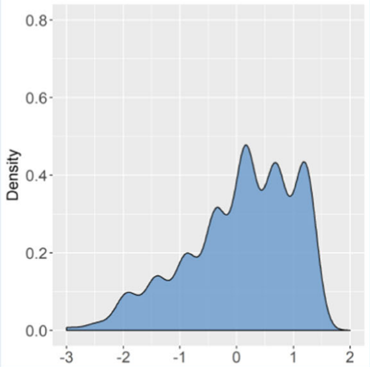
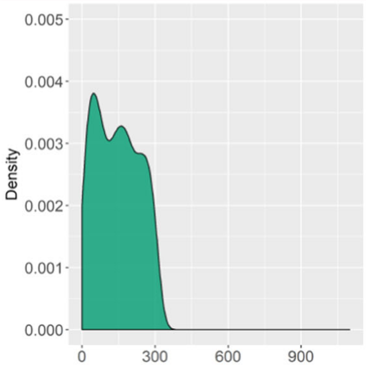
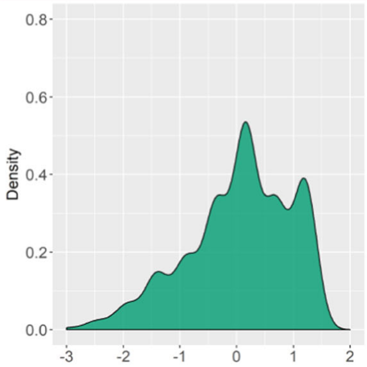
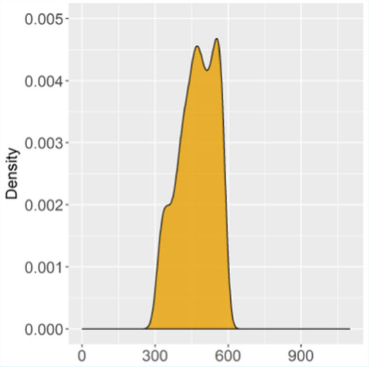
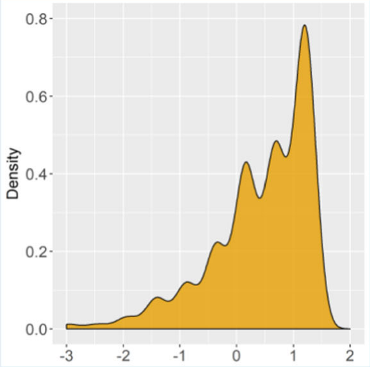
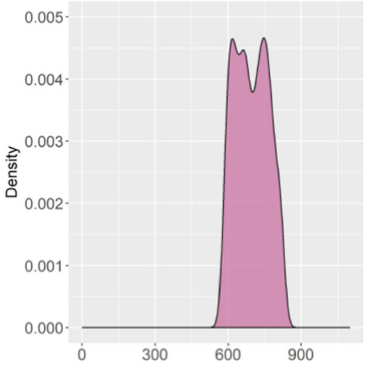
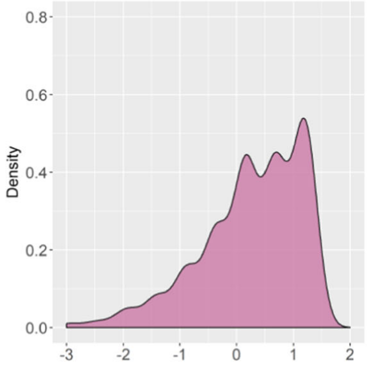
A deeper exploration of the salience of observed features as discriminants of the automatically identified computed phenotypes is illustrated on Table 8. For each pair of clusters and each observable variable, we can explicate the contrast specific distributions (for continuous) and factor-levels (for categorical) features. Demonstrating all possible scenarios is impractical, due to the number of possible combinations (of cluster pairs) and the large number of features. However, Table 8 illustrates this capability specifically for the "*onset_delta.x*" and "*trunk_max*" covariates. Graphical exploration of violin plots, box-and-whisker plots, histogram of estimated density plots provide a mechanism to examine the core univariate characteristics that segregate the derived computed phenotypes.

Dimensionality reduction using t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton 2008; Dinov 2018) provided an additional confirmation of the intriguing four computable phenotypes identified independently by the unsupervised clustering. Figure 4 illustrates the flat 2D t-SNE manifold representation of the data where each case is post-hoc color-coded by its corresponding derived cluster label. The clear separation of the computed phenotypes in the t-SNE space, with only minor label overlaps, provides evidence of structure in the multivariate data recovered independently by the clustering and the t-SNE embedding methods.

Discussion and Conclusions

Hothorn and colleagues (Hothorn and Jung 2014) applied conditional random forests to model the trajectory of ALSFRS score over time, and their results indicated high variability of the ALSFRS ratio among the study subjects (RMSE = 0.5208, Pearson correlation = 0.4014). Gomeni et al. (Gomeni et al. 2014) used a non-linear Weibull model to describe the ALS disease progression based on ALSFRS-R, and identified two clusters of trajectories (slow progression vs fast progression) using stepwise logistic regression. Taylor and co-workers (Taylor et al. 2016) selected 3742 out of

Table 8 Examples of deeper exploration explicating the relation between machine-derived computed phenotypes (clusters) and observable clinical phenotypes (in terms of some ALS variables)

Features Clusters	*onset_delta.x*	(normalized) *trunk_max*
Cluster 1		
Cluster 2		
Cluster 3		
Cluster 4		

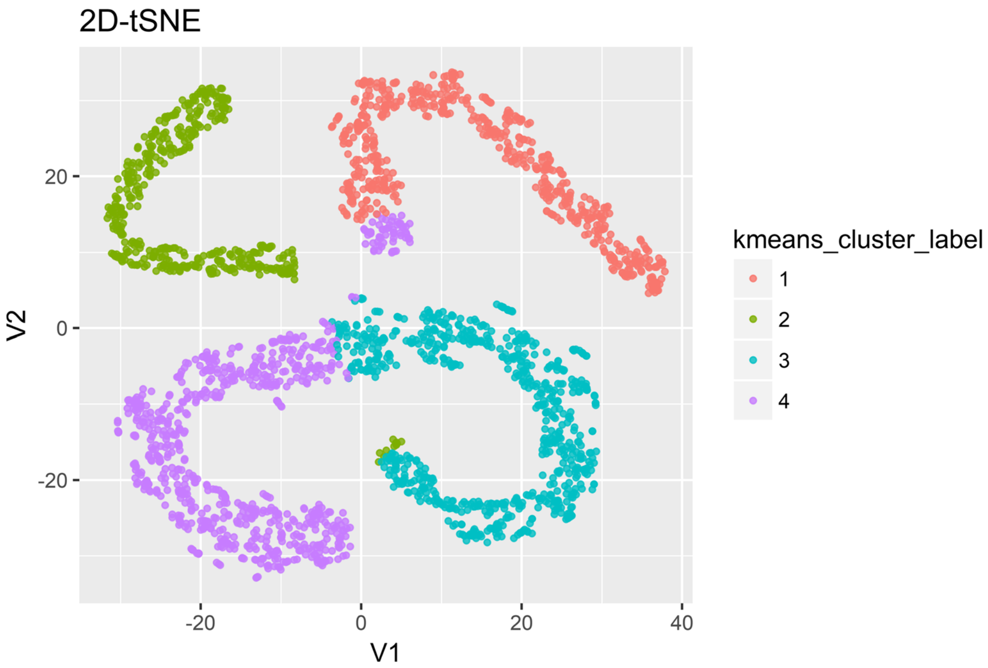


Fig. 4 2D t-SNE manifold representation of the data where each case is post-hoc color-coded by its corresponding derived cluster label

10,700 patients, using completeness in ALSFRS-R and forced vital capacity (FVC) records, to report a random forest model outperforming other approaches in predicting 6-month ALSFRS-R score ($R^2 = 0.582$, RMSE = 0.470). Ong et al. (Ong et al. 2017) found the exponential model was the best fit and categorized the subjects into fast and slow progressors based on score declined per day. Beaulieu-Jones and colleagues (Beaulieu-Jones and Moore 2017) implemented different imputation methods showing imputation does not have a significant effect on prediction. They also discovered that the progression of ALS approximately follows normal distribution.

ALS studies are complicated by the heterogeneity of this neurodegenerative disorder as well as the enormous challenges in handling, processing, and interpreting massive amounts of multi-source incongruent data serving as proxy of the onset and advancement of the disease. Providing clinical decision support to accurately prognosticate the course of the disease would be extremely valuable to identify reliable individualized predictions of the likely progression of ALS. In this study, we examined model-based (e.g., linear models) and model-free (e.g., machine-learning) methods to predict clinically relevant outcomes, e.g., change of ALSFRS score of FVC. Using independent training data (for model estimation and machine learning) and testing data (for assessing the accuracy and reliability of the predictions) we quantified the performance and compared the results of different forecasting techniques. The change of the Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) slope between 3 and 12 month of monitoring was used as a diagnostic predictive outcome variable.

A 2018 postmortem study of ALS heterogeneity identified factors contributing to survival duration (0–10 years) in 800 patients (Pfohl et al. 2018). Using survival regression, classification models (GML and RF), and dimensionality reduction the authors identified 38 salient measures (e.g., respiratory function, FVC, oxygen saturation, negative inspiratory force) that predicted ALSFRS-R, and other clinical outcomes. The main conclusion reported by the authors suggests that for personalized forecasting of survival, new independent metrics, as well as revisions to current metrics (e.g., ALSFRS-R), may need to be developed to capture the observed ALS heterogeneity.

Our study was based on training and testing data of patients and controls from the open Pooled Resource Open-Access ALS Clinical Trials Database (PRO-ACT) archive. The results of the study included both positive and negative findings about the impact of alternative processing protocols and the power of different analytical methods to forecast specific clinical phenotypes. The best model-based and model-free results highlight that initial measures of ALSFRS and its subquestions most significantly contributed to the overall prediction the ALSFRS at 1 year (12-month follow up). Other variables that contributed to overall prediction included time since disease onset and various measures of hematocrit, creatinine, potassium, calcium, and alanine transaminase (ALT). The contribution of creatinine confirms some prior reports, however, in this study, uric acid was not a strong clinical predictor. That changes in ALSFRS over the initial 3 months of data collection are the most significant predictors of future ALSFRS score, which is not surprising given the curvilinear nature of this clinical measure. Our findings highlight that the

ALSFRS is an informative measure, but new biologic markers are needed to improve univariate disease progression prediction and increase the reliability of such machine-learning based decision support systems. It is worth pointing out that strictly speaking, the values of each single ALSFRS item is intrinsically a discrete (ordinal/categorical) covariate. However, most studies model ALSFRS scores as continuous variables. From a methodological point of view, depending on the specific analytical strategy, treating ALSFRS as a continuous outcome may or may not be valid, e.g., classification and clustering of categorical outcomes is meaningful, but continuous regression linear modeling may be less appropriate.

Albeit direct implementations of these findings in clinical settings may still be in the future, our technique provides a roadmap for scientifically rigorous variable selection, computationally efficient inference using model-based and model-free methods, reproducible scientific discovery, and collaborative validation.

We also determined that the data preprocessing protocols may impact the results of the predictive analytics. We tried several alternative data preprocessing strategies. The main problems with preserving the bulk of the information content of the data related to data incongruency (e.g., different sampling rates) and data heterogeneity (e.g., missingness). To address the longitudinal incongruency within one covariate and one subject, we used the arithmetic average of all observed values. Initially, to handle missing values, we employed multiple imputation across all data elements and subjects. However, shot-gun approaches did not work well. This may be due to violations of the missing-at-random assumption or instability of the iterative multiple imputation algorithm on this large-scale data. Our final strategy for managing the missing data included a blended approach of (1) selective triaging of cases and variables, (2) imputing missing cells by the median value of a feature, and (3) multiple imputation on selected cases and data elements.

The Synapse ALS challenge initiative presented cluster analysis as an important approach in examining and simulating ALS disease progression. Our classification and prediction of specific univariate clinical traits were not impressive. It's possible that better classification models can be built, estimated, and validated on larger cohorts of patients sharing similar (congruent) clinical features. Our classification attempts were not reliable in identifying specific univariate clinically-relevant outcomes. There is also little evidence of prior successful classification results in other peer-reviewed publications. However, our unsupervised clustering analysis generated highly reliable, reproducible, and consistent computable phenotypes representing four complementary sub-cohorts, which are discriminated by multivariate clinical factors. This salient phenotypic clustering based on families of clinical features may explain the observed heterogeneity of ALS morbidity.

Our investigation faced two difficult challenges. The first one was transforming and harmonizing the longitudinal data into signature vectors congruent across subjects including maximum, minimum, median and the rate. Other investigators have suggested using the "first visit", "the last visit", "mean" and "standard deviation" to summarize the longitudinal information. However, bringing additional per-encounter information, does not resolve the sampling incongruency. Our approach compromises between increasing the volume of preserved information and minimizing the sampling incongruency. There is some evidence that ALSFRS is not linear but curvilinear (Gordon et al. 2010), however, further investigation is warranted. Another interesting prospective examination may include information about adverse events. It's not clear if utilizing observed adverse events may improve the predictive power of the algorithms to discriminate between different ALS clinical phenotypes.

The same compromise also applies when dealing with the missing data. On the one hand, to preserve the amount of observed information, we need to keep as many features as possible. On the other hand, a large number of data elements increases the demand on the subsequent data harmonization and aggregation. Larger number of subjects included in the training and testing collections would increase the predictive power of the methods (using the training data) and improve the forecasting results (using the testing data). The strategy we used was first triaging highly missing subjects and features, followed by handling missing testing data to ensure the final training-data-derived predictors are applicable to the validation data. This homogenized the predictive features and allowed aggregation of the data that led to consistent and reliable clustering results. When some variables are highly correlated, the multiple imputation may leave a few elements as missing, e.g., when the imputation algorithm does not converge. Novel numerical methods may help to alleviate the convergence problem. Adding an observation-derived prior could shrink the covariance matrix and ease the difficulty in converging. In this study, 1% of the total observations are used for constructing the empirical prior.

This study has limitations. The PRO-ACT data contains mostly subjects that entered a clinical trial. These subjects may not reflect the true population of ALS subjects given study-specific exclusion criteria. For example, subjects with dysarthria or dysphagia or early onset respiratory insufficiency may be ineligible for a trial and may not be represented in this dataset. Nonetheless, the presented study does offer important insights into modeling complex ALS data. This is important especially when considering designing prospective ALS studies or deciding on data elements to collect in large ongoing clinical studies. Our choice to only use the original ALSFRS scores may be an additional limitation as the revised version (ALSFRS-R) includes additional assessments, e.g., dyspnea, orthopnea and ventilatory support, which could potentially provide complementary information for forecasting ALS progression or identifying ALS clinical phenotypes.

There are some reports that ALSFRS may be time sensitive (Pfohl et al. 2018), which would make it a tricky clinical outcome to track the longitudinal progression of ALS. Our unsupervised approach is less sensitive to the dynamics of specific clinical outcomes, as we make no assumptions on a univariate response, but rather consider all observables as time-varying predictors. In this manuscript, we did not explore specifically predicting survival duration, however, elsewhere, we have demonstrated alternative non-parametric strategies to forecast patient survival (Huang et al. 2017).

Based on the reported findings, a deductive approach examining the specific from the general may be employed to develop clinical decision support system for prospective (individualized) personalized medicine based on first understanding the (global) ALS population characteristics. Due to low signal relative to normal biological and medical variation, accurate and highly-robust predictions of ALSFRS/ALSFRS-R may not be possible using the type of data provided in the PRO-ACT archive. For instance, a Rasch-analysis of the dimensionality, reliability and validity of ALSFRS-R suggested the need for improvements of its metric quality (Franchignoni et al. 2013). The three factors identified by the authors included bulbar function, fine and gross motor function, and respiratory function.

Information Sharing Statement

Following open-science principles, we share the entire data processing protocol to enable community-wide validation, improvements, and collaborative transdisciplinary research on complex healthcare and biomedical challenges. To ensure scientific reproducibility and promote community validation of our methods and findings, the R-based ALS Predictive Analytics source-code is released under permissive LGPL license on our GitHub repository (https://github.com/SOCR/ALS_PA).

Acknowledgements Colleagues from the Statistics Online Computational Resource (SOCR), Center for Complexity and Self-management of Chronic Disease (CSCD), Big Data Discovery Science (BDDS), and the Michigan Institute for Data Science (MIDAS) provided constructive feedback about this study.

Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: Neurological Clinical Research Institute, MGH; Northeast ALS Consortium; Novartis; Prize4Life Israel; Regeneron Pharmaceuticals, Inc.; Sanofi; Teva Pharmaceutical Industries, Ltd.

Finally, the authors are deeply indebted to the journal editors and the anonymous reviewers who provided valuable recommendations and constructive critiques that improved the manuscript.

Author's Contributions MT: developed techniques, conducted analyses, and wrote manuscript.

CG: developed techniques, conducted analyses, and wrote manuscript.

SAG: conceptualized the study and wrote manuscript.

AK: informatics, data analytics, and wrote manuscript.

BM: biostatistical methodology and wrote manuscript.

YG: conducted analyses, and wrote manuscript.

IDD: conceptualized the study, developed methods, conducted analyses, and wrote manuscript.

Funding This research was partially supported by NSF grants 1734853, 1636840, 1416953, 0716055 and 1023115, NIH grants P20 NR015331, P50 NS091856, UL1TR002240, P30 DK089503, U54 EB020406, P30 AG053760, and K23 ES027221, and the Elsie Andresen Fiske Research Fund. These funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Compliance with Ethical Standards

Ethics Approval and Consent to Participate University of Michigan Institutional Review Board (IRB) approval (HUM00115107) was obtained prior to managing, processing and analyzing the PRO-ACT data.

Competing Interests S.A.G. Dr. Goutman has received research support from the NIH/NIEHS (K23ES027221), Agency for Toxic Substances and Disease Registry/Centers for Disease Control, the ALS Association, Target ALS, Cytokinetics, and Neuralstem, Inc., and consulted for Cytokinetics.

References

Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 57*(3), 273–291.

Allen-Zhu, Z., & Hazan, E. (2016). Variance reduction for faster non-convex optimization. *in International Conference on Machine Learning*.

Atassi, N., Berry, J., Shui, A., Zach, N., Sherman, A., Sinani, E., Walker, J., Katsovskiy, I., Schoenfeld, D., Cudkowicz, M., & Leitner, M. (2014). The PRO-ACT database design, initial analyses, and predictive features. *Neurology, 83*(19), 1719–1725.

Beaulieu-Jones, B.K., & Moore, J.H. (2017). Missing data imputation in the electronic health record using deeply learned autoencoders, in *Pacific Symposium on Biocomputing 2017*, R.B. Altman, et al., Editors. p. 207–218.

Bergsma, W., Croon, M.A., & Hagenaars, J.A. (2009). *Marginal models: For dependent, clustered, and longitudinal categorical data*. Springer Science & Business Media.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning, 8*(3–4), 231–357.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of statistical software, 45*(3).

Carreiro, A. V., Amaral, P. M. T., Pinto, S., Tomás, P., de Carvalho, M., & Madeira, S. C. (2015). Prognostic models based on patient snapshots and time windows: Predicting disease progression to assisted ventilation in amyotrophic lateral sclerosis. *Journal of biomedical informatics, 58*, 133–144.

Cedarbaum, J. M., & Stambler, N. (1997). Performance of the amyotrophic lateral sclerosis functional rating scale (ALSFRS) in multicenter clinical trials. *Journal of the Neurological Sciences, 152*, s1–s9.

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., & Nakanishi, A. (1999). The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the neurological sciences, 169*(1), 13–21.

Chatterjee, S., & Hadi, A.S. (2015). *Regression analysis by example*. John Wiley & Sons.

De Sa, J.M. (2012). *Pattern recognition: concepts, methods and applications*. Springer Science & Business Media.

Dinov, I. D. (2016). Volume and value of big healthcare data. *Journal of Medical Statistics and Informatics, 4*(1), 1–7.

Dinov, I. D. (2018). Data science and predictive analytics: Biomedical and health applications using R, Springer, *Computer Science*, https://doi.org/10.1007/978-3-319-72347-1.

Dinov, I. D., Heavner, B., Tang, M., Glusman, G., Chard, K., Darcy, M., Madduri, R., Pa, J., Spino, C., Kesselman, C., Foster, I., Deutsch, E. W., Price, N. D., van Horn, J. D., Ames, J., Clark, K., Hood, L., Hampstead, B. M., Dauer, W., & Toga, A. W. (2016). Predictive big data analytics: A study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS One, 11*(8), e0157077.

Edwards, N., Wu, X., & Tseng, C.-W. (2009). An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra. *Clinical Proteomics, 5*(1), 23–36.

Fiedler, M., et al. (2006). *Linear optimization problems with inexact data*. Springer Science & Business Media.

Filzmoser, P., Baumgartner, R., & Moser, E. (1999). A hierarchical clustering method for analyzing functional MR images. *Magnetic Resonance Imaging, 17*(6), 817–826.

Franchignoni, F., Mora, G., Giordano, A., Volanti, P., & Chiò, A. (2013). Evidence of multidimensionality in the ALSFRS-R scale: A critical appraisal on its measurement properties using Rasch analysis. *Journal of Neurology, Neurosurgery, and Psychiatry, 84*(12), 1340–1345.

Gomeni, R., Fava, M., & P.R.O.-A.A.C.T. Consortium. (2014). Amyotrophic lateral sclerosis disease progression model. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 15*(1–2), 119–129.

Gong, P., et al. (2013). A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. in *International Conference on Machine Learning*.

Gordon, P. H., Cheng, B., Salachas, F., Pradat, P. F., Bruneteau, G., Corcia, P., Lacomblez, L., & Meininger, V. (2010). Progression in ALS is not linear but is curvilinear. *Journal of Neurology, 257*(10), 1713–1717.

Grigull, L., et al. (2016). Diagnostic support for selected neuromuscular diseases using answer-pattern recognition and data mining techniques: A proof of concept multicenter prospective trial. *BMC Medical Informatics and Decision Making, 16*(1), 1.

Hothorn, T., & Jung, H. H. (2014). RandomForest4Life: A random Forest for predicting ALS disease progression. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 15*(5–6), 444–452.

Huang, Z., Zhang, H., Boss, J., Goutman, S. A., Mukherjee, B., Dinov, I. D., Guan, Y., & for the Pooled Resource Open-Access ALS Clinical Trials Consortium. (2017). Complete hazard ranking to analyze right-censored data: An ALS survival study. *PLOS Computational Biology, 13*(12), e1005887.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition letters, 31*(8), 651–666.

Jain, P., & Kar, P. (2017). Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning, 10*(3–4), 142–336.

Kai-Hsiang, C., et al. (1999). Model-free functional MRI analysis using Kohonen clustering neural network and fuzzy C-means. *IEEE Transactions on Medical Imaging, 18*(12), 1117–1128.

Kuffner, R., et al. (2015). Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nature Biotechnology, 33*(1), 51–57.

Maaten, L.v.d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(Nov), 2579–2605.

Mairal, J. (2015). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization, 25*(2), 829–855.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., Poewe, W., Mollenhauer, B., Klinik, P. E., Sherer, T., Frasier, M., Meunier, C., Rudolph, A., Casaceli, C., Seibyl, J., Mendick, S., Schuff, N., Zhang, Y., Toga, A., Crawford, K., Ansbach, A., de Blasio, P., Piovella, M., Trojanowski, J., Shaw, L., Singleton, A., Hawkins, K., Eberling, J., Brooks, D., Russell, D., Leary, L., Factor, S., Sommerfeld, B., Hogarth, P., Pighetti, E., Williams, K., Standaert, D., Guthrie, S., Hauser, R., Delgado, H., Jankovic, J., Hunter, C., Stern, M., Tran, B., Leverenz, J., Baca, M., Frank, S., Thomas, C. A., Richard, I., Deeley, C., Rees, L., Sprenger, F., Lang, E., Shill, H., Obradov, S., Fernandez, H., Winters, A., Berg, D., Gauss, K., Galasko, D., Fontaine, D., Mari, Z., Gerstenhaber, M., Brooks, D., Malloy, S., Barone, P., Longo, K., Comery, T., Ravina, B., Grachev, I., Gallagher, K., Collins, M., Widnell, K. L., Ostrowizki, S., Fontoura, P., Ho, T., Luthman, J., Brug, M. . ., Reith, A. D., & Taylor, P. (2011). The Parkinson progression marker initiative (PPMI). *Progress in Neurobiology, 95*(4), 629–635.

Markus, K. A. (2012). Principles and practice of structural equation modeling by Rex B. Kline. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(3), 509–512.

Moon, S. W., et al. (2015a). Structural neuroimaging genetics interactions in Alzheimer’s disease. *Journal of Alzheimer's Disease, 48*(4), 1051–1063.

Moon, S. W., Dinov, I. D., Hobel, S., Zamanyan, A., Choi, Y. C., Shi, R., Thompson, P. M., Toga, A. W., & for the Alzheimer's Disease Neuroimaging Initiative. (2015b). Structural brain changes in early-onset Alzheimer's disease subjects using the LONI pipeline environment. *Journal of Neuroimaging, 25*(5), 728–737.

Ong, M.-L., Tan, P. F., & Holbrook, J. D. (2017). Predicting functional decline and survival in amyotrophic lateral sclerosis. *PLoS One, 12*(4), e0174925.

Pfohl, S. R., Kim, R. B., Coan, G. S., & Mitchell, C. S. (2018). Unraveling the complexity of amyotrophic lateral sclerosis survival prediction. *Frontiers in Neuroinformatics, 12*(36).

Rodriguez-Galiano, V., et al. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing, 67*, 93–104.

Saitta, S., Kripakaran, P., Raphael, B., & Smith, I. F. C. (2010). Feature selection using stochastic search: An application to system identification. *Journal of Computing in Civil Engineering, 24*(1), 3–10.

Saykin, A. J., Shen, L., Yao, X., Kim, S., Nho, K., Risacher, S. L., Ramanan, V. K., Foroud, T. M., Faber, K. M., Sarwar, N., Munsie, L. M., Hu, X., Soares, H. D., Potkin, S. G., Thompson, P. M., Kauwe, J. S., Kaddurah-Daouk, R., Green, R. C., Toga, A. W., Weiner, M. W., & Alzheimer's Disease Neuroimaging Initiative. (2015). Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. *Alzheimers & Dementia, 11*(7), 792–814.

Steinberg, D., & Colla, P. (2009). Cart: classification and regression trees. *The Top Ten Algorithms in Data Mining, 9*, 179.

Su, Y.-S., et al. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software, 45*(2), 1–31.

Tamás Kincses, Z., Johansen-Berg, H., Tomassini, V., Bosnell, R., Matthews, P. M., & Beckmann, C. F. (2008). Model-free characterization of brain functional networks for motor sequence learning using fMRI. *NeuroImage, 39*(4), 1950–1958.

Taylor, A. A., Fournier, C., Polak, M., Wang, L., Zach, N., Keymer, M., Glass, J. D., Ennist, D. L., & The Pooled Resource Open-Access ALS Clinical Trials Consortium. (2016). Predicting disease progression in amyotrophic lateral sclerosis. *Annals of Clinical and Translational Neurology, 3*(11), 866–875.

Westeneng, H.-J., Debray, T. P. A., Visser, A. E., van Eijk, R. P. A., Rooney, J. P. K., Calvo, A., Martin, S., McDermott, C. J., Thompson, A. G., Pinto, S., Kobeleva, X., Rosenbohm, A., Stubendorff, B., Sommer, H., Middelkoop, B. M., Dekker, A. M., van Vugt, J. J. F. A., van Rheenen, W., Vajda, A., Heverin, M., Kazoka, M., Hollinger, H., Gromicho, M., Körner, S., Ringer, T. M., Rödiger, A., Gunkel, A., Shaw, C. E., Bredenoord, A. L., van Es, M. A., Corcia, P., Couratier, P., Weber, M., Grosskreutz, J., Ludolph, A. C., Petri, S., de Carvalho, M., van Damme, P., Talbot, K., Turner, M. R., Shaw, P. J., al-Chalabi, A., Chiò, A., Hardiman, O., Moons, K. G. M., Veldink, J. H., & van den Berg, L. H. (2018). Prognosis for patients with amyotrophic lateral sclerosis: Development and validation of a personalised prediction model. *The Lancet Neurology, 17*(5), 423–433.

Wismüller, A., Meyer-Bäse, A., Lange, O., Auer, D., Reiser, M. F., & Sumners, D. W. (2004). Model-free functional MRI analysis based on unsupervised clustering. *Journal of Biomedical Informatics, 37*(1), 10–18.

Wistuba, M., Schilling, N., & Schmidt-Thieme, L.. (2015). Sequential model-free Hyperparameter tuning. in Data mining (ICDM), 2015 IEEE International Conference on.

Witten, I.H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Zach, N., Ennist, D. L., Taylor, A. A., Alon, H., Sherman, A., Kueffner, R., Walker, J., Sinani, E., Katsovskiy, I., Cudkowicz, M., & Leitner, M. L. (2015). Being PRO-ACTive: What can a clinical trial database reveal about ALS? *Neurotherapeutics, 12*(2), 417–423.

Zhang, G. P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30*(4), 451–462.

Affiliations

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, University of Michigan, Ann Arbor, MI 48109, USA

[2] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

[3] Department of Neurology, University of Michigan, Ann Arbor, MI 48109, USA

[4] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

[5] Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, USA