Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study

Faraz Faghri, Fabian Brunn, Anant Dadu, PARALS consortium, ERRALS consortium*, Elisabetta Zucchi, Ilaria Martinelli, Letizia Mazzini, Rosario Vasta, Antonio Canosa, Cristina Moglia, Andrea Calvo, Michael A Nalls, Roy H Campbell, Jessica Mandrioli†, Bryan J Traynor†, Adriano Chiò†*

Summary

Background Amyotrophic lateral sclerosis (ALS) is known to represent a collection of overlapping syndromes. Various classification systems based on empirical observations have been proposed, but it is unclear to what extent they reflect ALS population substructures. We aimed to use machine-learning techniques to identify the number and nature of ALS subtypes to obtain a better understanding of this heterogeneity, enhance our understanding of the disease, and improve clinical care.

Methods In this retrospective study, we applied unsupervised Uniform Manifold Approximation and Projection [UMAP]) modelling, semi-supervised (neural network UMAP) modelling, and supervised (ensemble learning based on LightGBM) modelling to a population-based discovery cohort of patients who were diagnosed with ALS while living in the Piedmont and Valle d'Aosta regions of Italy, for whom detailed clinical data, such as age at symptom onset, were available. We excluded patients with missing Revised ALS Functional Rating Scale (ALSFRS-R) feature values from the unsupervised and semi-supervised steps. We replicated our findings in an independent population-based cohort of patients who were diagnosed with ALS while living in the Emilia Romagna region of Italy.

Findings Between Jan 1, 1995, and Dec 31, 2015, 2858 patients were entered in the discovery cohort. After excluding 497 (17%) patients with missing ALSFRS-R feature values, data for 42 clinical features across 2361 (83%) patients were available for the unsupervised and semi-supervised analysis. We found that semi-supervised machine learning produced the optimum clustering of the patients with ALS. These clusters roughly corresponded to the six clinical subtypes defined by the Chiò classification system (ie, bulbar, respiratory, flail arm, classical, pyramidal, and flail leg ALS). Between Jan 1, 2009, and March 1, 2018, 1097 patients were entered in the replication cohort. After excluding 108 (10%) patients with missing ALSFRS-R feature values, data for 42 clinical features across 989 patients were available for the unsupervised and semi-supervised analysis. All 1097 patients were included in the supervised analysis. The same clusters were identified in the replication cohort. By contrast, other ALS classification schemes, such as the El Escorial categories, Milano-Torino clinical staging, and King's clinical stages, did not adequately label the clusters. Supervised learning identified 11 clinical parameters that predicted ALS clinical subtypes with high accuracy (area under the curve 0·982 [95% CI 0·980–0·983]).

Interpretation Our data-driven study provides insight into the ALS population substructure and confirms that the Chiò classification system successfully identifies ALS subtypes. Additional validation is required to determine the accuracy and clinical use of these algorithms in assigning clinical subtypes. Nevertheless, our algorithms offer a broad insight into the clinical heterogeneity of ALS and help to determine the actual subtypes of disease that exist within this fatal neurodegenerative syndrome. The systematic identification of ALS subtypes will improve clinical care and clinical trial design.

Funding US National Institute on Aging, US National Institutes of Health, Italian Ministry of Health, European Commission, University of Torino Rita Levi Montalcini Department of Neurosciences, Emilia Romagna Regional Health Authority, and Italian Ministry of Education, University, and Research.

Introduction

Amyotrophic lateral sclerosis (ALS) is one of the most common forms of neurodegeneration, accounting for approximately 6000 deaths in the USA and 11 000 deaths in Europe, annually.[1] ALS is characterised by progressive paralysis of limb and bulbar musculature, and typically leads to death within 3–5 years of symptom onset. Medications only minimally slow the rate of progression, so treatment focuses on symptomatic management.

Lancet Digit Health 2022

Published Online March 24, 2022 https://doi.org/10.1016/S2589-7500(21)00274-0

For the Italian translation of the abstract see Online for appendix 1

For the German translation of the abstract see Online for appendix 2

*Consortia members are listed in appendix 3 (pp 24–25)

†Contributed equally

Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, US National Institute on Aging, Bethesda, MD, USA (B J Traynor MD, F Faghri PhD); **Center for Alzheimer's and Related Dementias, US National Institute on Aging, Bethesda, MD, USA** (F Faghri, M A Nalls PhD, A Dadu BS); **Data Tecnica International, Glen Echo, MD, USA** (F Faghri, A Dadu, M A Nalls); **Department of Computer Science, University of Illinois at Urbana–Champaign, Champaign, IL,USA** (F Faghri, F Brunn MS, A Dadu, Prof R H Campbell PhD); **Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy** (E Zucchi MD, J Mandrioli MD); **Neurology Unit, Department of Neurosciences, Azienda Ospedaliero Universitaria di Modena, Modena, Italy** (I Martinelli MD, J Mandrioli); **ALS Centre, Department of Neurology, Maggiore della Carità University Hospital, Novara, Italy** (L Mazzini MD); **Rita Levi Montalcini, Department of Neuroscience, University of Turin, Turin, Italy** (R Vasta MD, A Canosa MD, C Moglia MD, A Calvo MD, Prof A Chiò MD); **Department of Neurology, Johns Hopkins University Medical Center,**

Baltimore, MD, USA (B J Traynor)**; Reta Lila Weston Institute, UCL Queen Square Institute of Neurology, University College London, London, UK** (B J Traynor)**; Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy** (Prof A Chiò)**; Neurology 1 and ALS Centre, Azienda Ospedaliero Universitaria Città della Salute e della Scienza, Turin, Italy** (Prof A Chiò)

Correspondence to: Dr Bryan J Traynor, Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892–3707, USA.
traynorb@mail.nih.gov

See Online for appendix 3

Research in context

Evidence before this study

We searched PubMed for articles published in English from database inception to Jan 5, 2021, about the use of machine learning and the identification of clinical subtypes within the amyotrophic lateral sclerosis (ALS) population, using the search terms "machine learning" AND "classification" AND "amyotrophic lateral sclerosis". The search identified 29 studies. Most of these studies used machine learning to diagnose ALS (on the basis of gait, imaging, electromyography, gene expression, proteomic, and metabolomic data) or to improve brain–computer interfaces. One study used machine-learning algorithms to stratify ALS post-mortem cortex samples into molecular subtypes on the basis of transcriptome data. A 2015 study crowdsourced the development of machine-learning algorithms to approximately 30 teams to try to obtain a consensus to identify subpopulations of patients with ALS. Although four categories of patients with ALS were identified, the clinical relevance of this approach was unclear, because all patients with ALS necessarily pass through an early and late stage of the disease. Furthermore, no attempt was made to discern which of the existing clinical classification systems (eg, the El Escorial criteria, the Chiò classification system, and the King's clinical staging system) can identify ALS subtypes.

ALS subtype identification has been explored using *t*-distributed stochastic neighbour embedding, and Uniform Manifold Approximation and Projection (UMAP) has also been used in the context of stratifying patients with ALS in two papers. Prognosis outcome and patient stratification have been modelled in a classification context using either real-life data or Pooled Resource Open-Access ALS Clinical Trials data. The Piedmont and Valle d'Aosta Registry for ALS (PARALS) data were also used for stratification of patients with ALS but most of the data in that study were not population-based. Our semi-supervised approach, based on a neural network and UMAP, is similar to work published by Sainburg and colleagues.

We concluded that there remained an unmet need to identify the ALS population substructure in a data-driven, non-empirical manner. Building on this conclusion, there was a need for a tool that reliably predicted the clinical subtype of patients with ALS. This knowledge would improve understanding of the clinical heterogeneity associated with this fatal neurodegenerative disease.

Added value of this study

This study developed a machine-learning algorithm to detect clinical subtypes of patients with ALS using clinical data collected from the 2858 Italian patients with ALS. Ascertainment of such patients within the catchment area was near complete, meaning that the dataset truly represented the ALS population. We replicated our approach using clinical data obtained from an independent cohort of 1097 Italian patients with ALS that had also been collected in a population-based, longitudinal manner. Semi-supervised learning based on UMAP applied to a multilayer perceptron neural network provided the optimum results based on visual inspection. The observed clusters equated to the six clinical ALS subtypes previously defined by the Chiò classification system (ie, bulbar, respiratory, flail arm, classical, pyramidal, and flail leg). Using a small number of clinical parameters, an ensemble-learning approach could predict the ALS clinical subtype with high accuracy (area under the curve 0·954).

Implications of all the available evidence

Additional validation is required to determine the accuracy and clinical use of these algorithms in assigning clinical subtypes. Nevertheless, our algorithms offer a broad insight into the clinical heterogeneity of ALS and help to determine the actual subtypes of disease that exist within this fatal neurodegenerative syndrome. The systematic identification of ALS subtypes could improve clinical care and clinical trial design.

Genetic advancements have shown that ALS is not a single entity and instead consists of a collection of syndromes in which the motor neurons degenerate. Alongside these multiple genetic aetiologies, there is broad variability in the disease's clinical manifestations, in terms of age at symptom onset, site of onset, rate and pattern of progression, and cognitive involvement. This clinical heterogeneity has hampered efforts to understand the cellular mechanisms underlying this fatal neurodegenerative syndrome and has hindered efforts to find effective therapies.

Given the importance of clinical heterogeneity within ALS, it is not surprising that there has been considerable effort over time to develop classification systems for patients. Examples include groupings based on family status,[2] clinical milestones,[3] neurophysiological measurements,[4] and diagnostic certainty.[5] Although useful, it is unclear whether any of these classification systems identify clinically meaningful subgroups within the ALS population, or merely represent human constructs based on empirical observations. Determining the correct number and nature of subgroups within the ALS population would be an important step towards understanding the disease. By extension, a reliable method to predict an individual patient's subgroup using data collected at the beginning of their illness would be helpful for clinical care and clinical trial design.

Our goal was to determine the disease subtypes existing within a deeply phenotyped, population-based collection of patients and to build predictor models to classify individuals according to their subtype using machine learning. The advantage of machine-learning approaches is their ability to identify complex relationships in a data-driven manner.

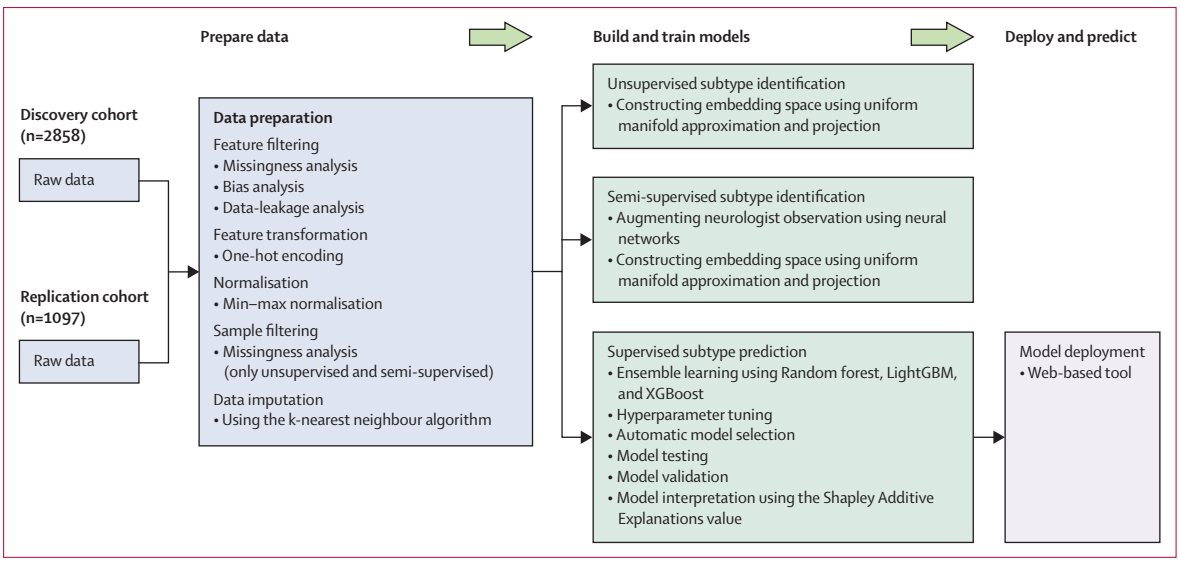

Figure 1: **Study workflow**
Unsupervised and semi-supervised machine learning were applied to clinical data collected from two population-based ALS registries (PARALS=2858 patients and ERRALS=1097 patients) to identify ALS clinical subtypes. Supervised machine learning was used to predict ALS subtypes on the basis of clinical parameters, and a web-based tool was built for clinical researchers to apply to their own data. ALS=amyotrophic lateral sclerosis.

Methods

Study design and participants

We explored the clinical subtypes of ALS by applying unsupervised and semi-supervised machine learning to deeply phenotyped, population-based cohorts of patients (see figure 1 for the analysis workflow). After identifying the ALS subtypes, we used supervised machine learning to build predictor models to classify individual patients.

The discovery cohort consisted of patients diagnosed with ALS while living in the Piedmont and Valle d'Aosta regions of Italy and entered in a population-based registry, known as the Piedmont and Valle d'Aosta Registry for ALS (PARALS; established Jan 1, 1995) during the study period.[6] This registry has near-complete case ascertainment within its catchment population of nearly 4·5 million inhabitants (appendix 3 p 1).[6]

To validate our results, we replicated the identification of the ALS subtypes using an independent cohort. The replication cohort consisted of patients diagnosed with ALS and living in the Emilia Romagna region of Italy, and entered in a population-based registry, known as the Emilia Romagna Region registry for ALS (ERRALS; established Jan 1, 2008).[7] The ERRALS catchment area included 4·4 million inhabitants.[7]

None of the patients with ALS who were enrolled in ERRALS were enrolled in PARALS, and there were no exclusion criteria for the registries. We used the discovery (PARALS) cohort as a training dataset, and the replication (ERRALS) cohort as the replication dataset in our machine-learning analyses.

An important feature of these two studies[6,7] is real-time collection, by study authors who were experienced ALS neurologists, of detailed data about patients throughout their illness. The data collection methods were standardised across the two registries to facilitate comparisons. Each patient was evaluated according to published classification schema that included: the El Escorial classification system,[5] family status (sporadic *vs* familial disease),[2] the Milano-Torino clinical staging system,[8] and the King's staging system.[3] The El Escorial diagnostic criteria for ALS classify patients into categories reflecting different degrees of diagnostic certainty.[5] The Milano-Torino staging system captures the clinical milestones corresponding to the loss of independence and function in patients with ALS.[8] The King's staging system is based on disease burden, as measured by clinical involvement and feeding or respiratory failure, and classifies patients into five stages, with stage 1 representing symptom onset and stage 5 being death.[3] The Revised ALS Functional Rating Score (ALSFRS-R) scale[9] includes 12 questions that each has a score ranging from 0 (no function) to 4 (full function) and is used to measure disease progression; the first three questions (part 1) of this ordinal scale evaluate the bulbar function of the patient. Patients were given an ALSFRS-R score and were dichotomised according to whether or not they were a *C9orf72* gene carrier (the most common genetic cause of ALS). The PARALS and ERRALS studies were approved by the local ethics committees (appendix 3 p 2). We anonymised all records in accordance with the Italian Personal Data Protection Code, Containing Provisions to Adapt the National Legislation to General Data Protection Regulation (Regulation [EU] 2016/679).

Preprocessing of the clinical data

The clinical data (appendix 3 p 13) were filtered before analysis. Features with non-random missingness

(eg, cancer type), high sampling bias (eg, place of birth), and features that could introduce data leakage (eg, tracheostomy, and an initial diagnosis of primary lateral sclerosis) were omitted from the analyses (appendix 3 pp 11–12). For unsupervised and semi-supervised ALS subtype identification, patients with missing values in the ALSFRS-R[9] feature were also excluded (497 [17%] of 2858 patients in the discovery cohort and 108 [10%] of 1097 patients in the replication cohort). By contrast, patients with missing ALSFRS-R data were included in the supervised analysis, because the ensemble-learning methods used can handle missingness. Thus, the prediction modelling used data for 2858 patients in the discovery cohort and 1097 patients in the replication cohort. Categorical features were encoded to numerical values using the one-hot encoding[10] method. Min–max normalisation was applied to numerical features to preserve the relationships among the original data and ensure a zero-to-one range.[11]

Data imputation

After data filtering and preprocessing, the following features had residual missingness that was distributed randomly across 15–20% of patients: forced vital capacity percentage at diagnosis; body-mass index (BMI) at 2 years before illness; rate of decline of BMI per month since 2 years before illness; weight 2 years before illness; BMI at diagnosis; height at diagnosis; and weight at diagnosis. To account for this, we used the k-nearest neighbour imputation method with k=5 neighbours to preserve the clusters.[12] The discovery and replication cohorts were imputed independently.

For more on **GenoML** see https://genoml.com/

Unsupervised machine learning

After preparing the data for analysis as described above, we did unsupervised machine learning. We hypothesised that machine-learning approaches could identify the number and nature of ALS subtypes when applied to a large, well-characterised population cohort. The primary outcome measure of our analyses was a comparison of the ALS subtype clusters defined by the approaches to the six clinical subtypes (ie, bulbar, respiratory, flail arm, classical, pyramidal, and flail leg) assigned manually by neurologists applying the Chiò classification system.[13] The clinical subtypes assigned by the Chiò classification system were not entered into the unsupervised algorithms and were not used to construct the patient clusters.

First, we used an unsupervised clustering approach to identify ALS subtypes by applying Uniform Manifold Approximation and Projection (UMAP) to the processed data. UMAP is used for non-linear dimensionality reduction to produce a low-dimensional projection of the data with the closest possible equivalent fuzzy topological structure.[14] This approach preserves the local and global structures existing within the data, along with reproducible and meaningful clusters. As a comparison, we applied dimensionality reduction methods such as principal component analysis, independent component analysis, and non-negative matrix factorisation to the data.

Semi-supervised machine learning

To further refine the clusters identified by UMAP alone, we processed the data using a multilayer perceptron neural network consisting of five hidden layers with 200, 100, 50, 25, and 3 neurons (appendix 3 p 4).[15] The network was trained with the clinical-type-at-1-year outcome labels related to the Chiò schema, using a Softmax classifier (which squashes raw class scores into normalised positive values that sum to one). After training the network with ten-times cross-validation, we extracted the activations of the last hidden layer and used them as the input for the UMAP algorithm.[14] This approach reduced the dataset dimensions from 72 dimensions at the start of the process to three dimensions at the end.

Supervised subtype prediction

Next, we applied a supervised-learning approach, called ensemble learning, to develop predictive models forecasting the ALS clinical subtype of a patient solely on the basis of clinical data obtained at the first neurology visit. Ensemble learning combines multiple learning algorithms to generate a better predictive model than a single learning algorithm could.[16] For supervised machine learning, we used GenoML, an open-source automated machine-learning package developed by the current authors.[17] Within this package, ensemble learning was used to develop predictive models to forecast the ALS clinical subtype of a patient solely on the basis of clinical data obtained at their first neurology visit. The stacking ensembles of three supervised machine-learning algorithms (Random Forest version 0.24.2,[18] LightGBM version 3.2.1,[19] and XGBoost version 1.4.2[20]) were evaluated, and the ensemble model that performed best was selected (see appendix 3 pp 5–6 for model selection and hyperparameter tuning). Feature reduction was done using recursive elimination to decrease the number of parameters included in the model without sacrificing accuracy. Internal validation on the discovery cohort and external validation on the replication cohort were used to assess performance and determine the best algorithms and parameters to use in the model using the logloss metric (appendix 3 p 2). Model performance was evaluated on the basis of various metrics, including accuracy, area under the curve (AUC), area under the precision-recall curve (AUPRC), and logloss. We used the Shapley Additive Explanations (SHAP)[21] approach to evaluate each clinical feature's influence in ensemble learning. This approach is used in game theory and assigns an importance (ie, SHAP) value to each feature to determine a player's contribution to success.[21] SHAP enhance understanding by creating accurate explanations for each observation in a dataset and bolstering trust if

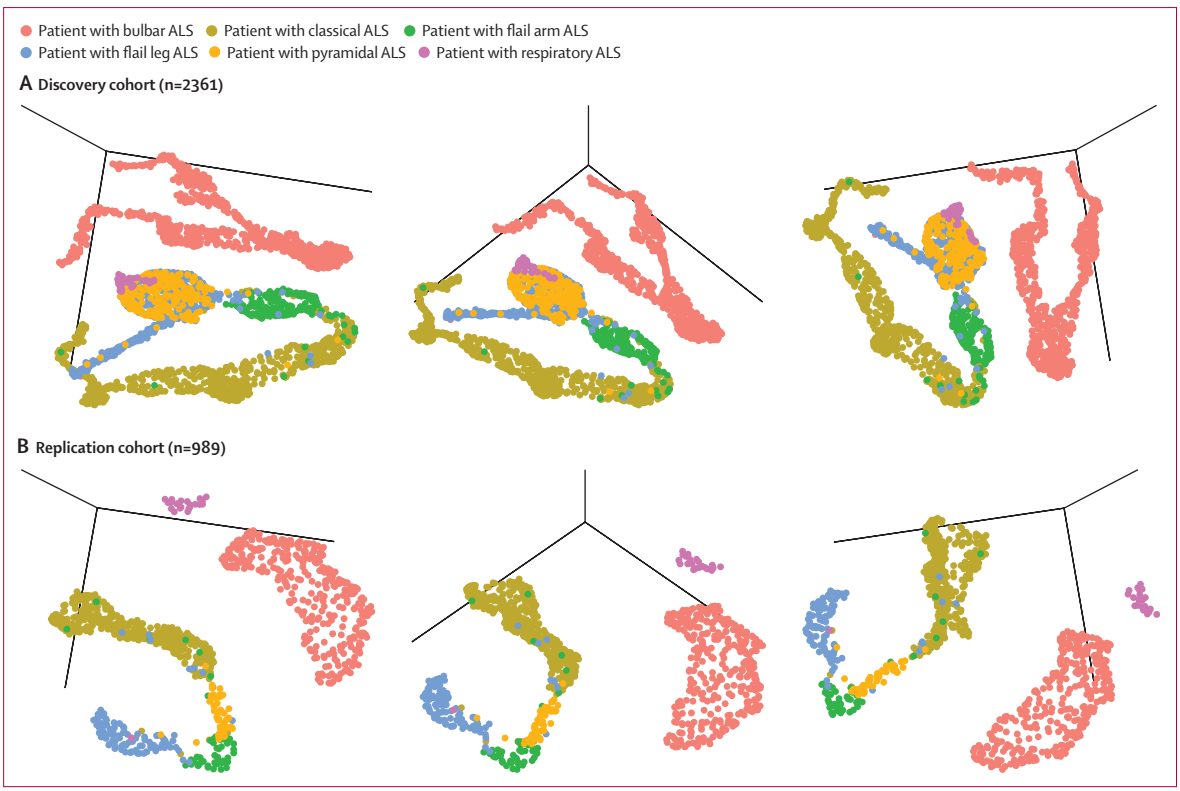

Figure 2: **The ALS subtypes identified by machine learning in the discovery and replication cohorts**
Three-dimensional projections for the discovery (ie, PARALS) cohort (A) and the replication (ie, ERRALS) cohort (B), with azimuthal rotations of 100° (left), 135° (centre), and 170° (right), which are symbolic of ALS subtypes as defined by the semi-supervised machine-learning algorithm that consisted of a uniform manifold approximation and projection algorithm applied to the output of a five-layer neural network. Colour coding using the Chiò classification system[13] was done after machine-learning cluster generation. Interactive three-dimensional graphs are available on the interactive Machine Learning for ALS website (https://share.streamlit.io/anant-dadu/machinelearningforals/main). ALS=amyotrophic lateral sclerosis.

the crucial variables for specific records conform to human domain knowledge and reasonable expectations. The interactive website was developed as an open-access, cloud-based platform to provide a simple-to-use tool that clinicians can access.

Computational tools and code availability

The data-analysis pipeline for this work was done in Python (version 3.6) using open-source libraries (NumPy [version 1.20.3], pandas [version 1.2.5], matplotlib [version 3.4.2], seaborn [version 0.11.1], plotly [version 4.14.2], scikit-learn [version 0.24.2], UMAP [version 0.5.0], XGBoost [version 1.4.2], LightGBM [version 3.2.1], GenoML [version 2v1.0.0b11], and TensorFlow [version 2.4.0]). We made our code publicly available to facilitate replication and future expansion of our work. Manuscript visualisations were created with tidyverse (version 1.3), ggplot2 (version 3.3.2), and plotly (version 4.9.2.2), and implemented in R (version 4.0.3). The exploratory data analysis was done with dlooker (version 0.5.4). The exploratory data analysis was the initial investigations done on data to discover any anomalies and to check assumptions with the help of summary statistics and graphical representations. The UpSet plot (also known as an attributes graph) was produced using UpSetR (version 1.4.0) software in R. UpSet plot analysis can only be done using complete data, whereas machine-learning analysis can be done and still be valid using samples from which missing ALSFRS data have been removed. The reporting guideline checklists are provided in appendix 3 (pp 26–31).

For the **interactive website** see https://share.streamlit.io/anant-dadu/machinelearningforals/main

Role of the funding source

The study sponsors had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Between Jan 1, 1995, and Dec 31, 2015, 2858 patients were entered in PARALS. The clinical and demographic details of this discovery cohort are given in appendix 3 (pp 13–16). The 66 clinical features collected for each patient are listed in appendix 3 (pp 11–12); for an exploratory data analysis describing the content of each feature see appendix 3 (pp 32–111). After filtering and excluding 497 (17%) patients who had missing values in the ALSFRS-R feature, data for 42 clinical features across 2361 (83%) of 2858 patients in the PARALS discovery cohort were available for the unsupervised analysis. We

For the **code** see https://github.com/ffaghri1/ALS-ML

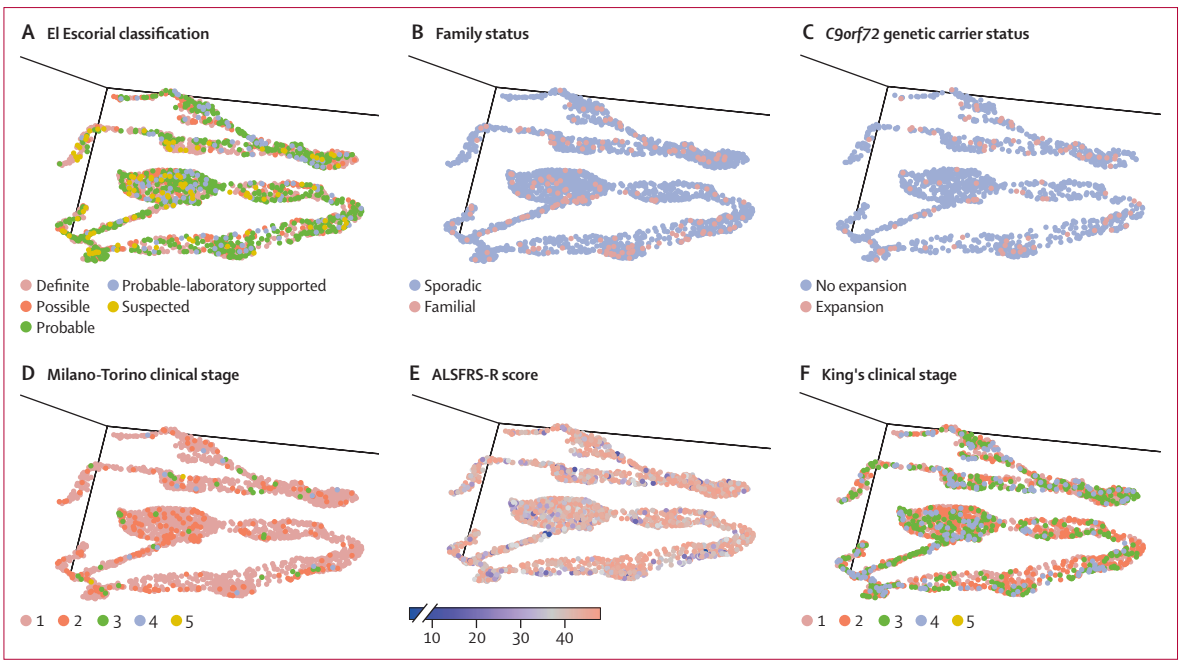

Figure 3: **Classification schema applied to the semi-supervised three-dimensional projection of the discovery (PARALS) cohort**
(A) The El Escorial classification system[5] assigns patients to five ALS categories on the basis of the extent of their disability. Laboratory supported means supported by neurophysiology, neuroimaging, and clinical laboratory tests. (B) Patients with a family history of ALS or sporadic disease. (C) Patients carrying the pathogenic repeat expansion mutation in *C9orf72*. (D) The Milano-Torino clinical staging classification system[8] assigns patients to stages 0–4 (minimal disability–most disability). (E) The ALSFRS-R score[9] rates a patient's physical function from 0 to 48 (most disability–no disability). (F) The King's clinical staging system[3] classifies patients into five stages from 1 (symptom onset) to 5 (death) according to the extent of their disability. ALS=amyotrophic lateral sclerosis. ALSFRS-R=Revised ALS Functional Rating Scale.[9]

included all 2858 (100%) patients in the semi-supervised analysis. Both the unsupervised and semi-supervised approaches identified multiple clusters of patients, representing distinct subtypes of ALS (for the results of the UMAP alone, see appendix 3 p 7; for the results of the neural network UMAP see figure 2A). Colour coding the patients according to the ALS clinical subtype assigned by a neurologist showed that the clusters roughly corresponded to the six clinical subtypes previously defined by the Chiò classification system[13] (primary outcome). Visually investigating these three-dimensional (3D) projections, the optimum separation of the patients into their clinical subtypes of ALS was obtained using the semi-supervised machine-learning approach. There was excellent discrimination of the bulbar, respiratory, flail arm, and classical subtypes of ALS. By contrast, the pyramidal and flail leg subtypes overlapped substantially although the flail leg variant did form a distinct tail that did not overlap with the other subtypes. Overall, we found that 787 (>99%) of 789 patients with bulbar, 42 (100%) of 42 patients with respiratory, 150 (91%) of 164 patients with flail arm, and 663 (94%) of 707 patients with classical ALS were assigned to the same subtype by both the neurologist and the semi-supervised algorithm.

For the replication study, between Jan 1, 2009, and March 1, 2018, 1097 patients were entered in ERRALS. For the unsupervised and semi-supervised analysis, we excluded 108 (10%) patients who had missing values in the ALSFRS-R feature; after filtering, data for 42 clinical features for 989 patients with ALS were available for analysis. We included all 1097 patients in the supervised analysis. The subtypes and clusters identified in the independent replication cohort are shown in figure 2B. Visually, the cluster pattern was similar to that observed in the discovery cohort, confirming the reproducibility of our data-driven approach. Interactive 3D graphs are available on the interactive Machine Learning for ALS website (see "Explore the ALS subtype topological space").

Our semi-supervised machine-learning algorithm was more accurate than the other dimensionality reduction approaches, such as principal component analysis and independent component analysis (appendix 3 p 8). Furthermore, other ALS classification schema, such as the El Escorial categories,[5] family status,[2] the presence or absence of the pathogenic *C9orf72* repeat expansion, Milano-Torino clinical staging,[8] ALSFRS-R score,[9] and King's clinical stages,[3] did not label the clusters in a meaningful, clinically useful manner (figure 3).

With the supervised (ensemble-learning) approach, if all available features (n=66) were included in the model, the clinical subtype of a patient was predicted with high accuracy (internal validation AUC 0·982 [95% CI

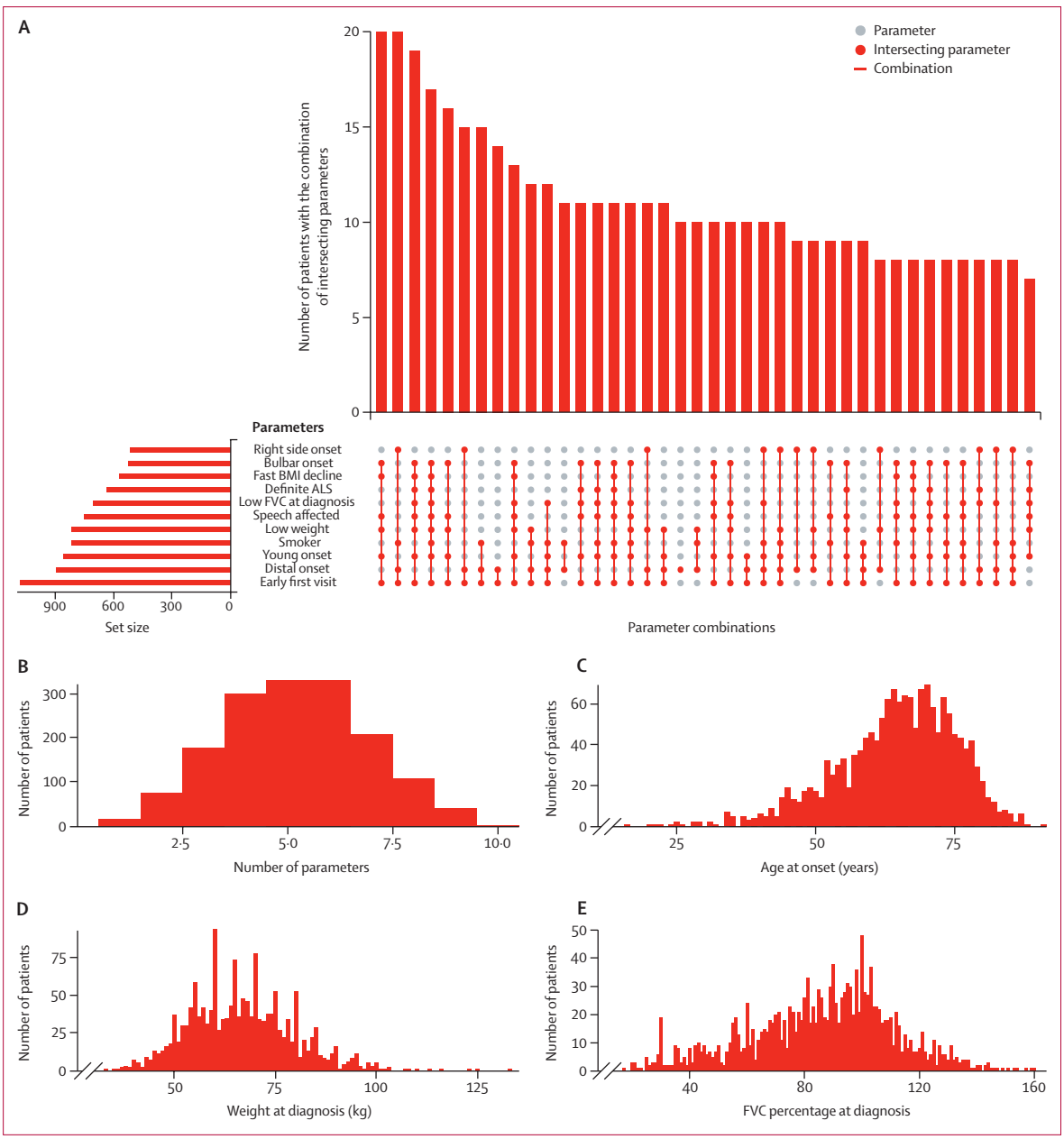

Figure 4: **UpSet plot of the clinical parameters used in the supervised machine-learning model to predict ALS clinical subtype**
Analysis was confined to 1584 ALS patients enrolled in PARALS with complete data and the figure was created using UpSetR software. Set size is the number of individuals with a specified parameter. (A) Graphical representation of the overlap between the 11 parameters that had the most substantial effects on the classification model. (B) Distribution of clinical parameters per patient (mean 5·1 [SD 1·7]). (C) Distribution of age at ALS onset. (D) Weight at diagnosis. (E) FVC percentage at diagnosis. ALS=amyotrophic lateral sclerosis. BMI=body-mass index. FVC=forced vital capacity.

0·979–0·984] and external validation AUC 0·954 [0·950–0·958]; appendix 3 pp 9, 17–23).

To increase the clinical utility of this approach, we used recursive feature elimination to decrease the number of parameters included in the model without sacrificing accuracy, and this reduced the number of parameters to 11. The full performance results for Accuracy, AUC, AUCPR, and logloss can be found in appendix 3 (p 17). The predictor model built with the top 11 factors was equally robust compared with the all-inclusive model (internal validation AUC 0·982 [95% CI 0·980–0·983] and external validation AUC 0·943 [0·939–0·947]; figure 4 and appendix 3 pp 9, 22–23), The table and figure 5 list the 11 parameters selected for the final model and their relative contributions to the model's precision. Finally, we implemented an interactive website that allows clinical researchers to determine the future clinical subtype of a patient with ALS on the basis of

	Relative importance to model precision	SD
Anatomical level at onset	1·000	0·078
Site of symptom onset	0·460	0·021
Onset side	0·132	0·023
Weight at diagnosis, kg	0·042	$4{\cdot}548\times10^{-4}$
El Escorial category[5] at diagnosis	0·033	0·003
ALSFRS-R part 1 score for speech	0·027	$8{\cdot}967\times10^{-4}$
Time from symptom onset to first ALSFRS-R measurement, days	0·020	0·000
Smoking status	0·019	$1{\cdot}980\times10^{-4}$
Age at symptom onset, years	0·015	0·000
Rate of body-mass index decline per month	0·014	0·000
Forced vital capacity percentage at diagnosis	0·013	0·000

ALSFRS-R=Revised Amyotrophic Lateral Sclerosis Functional Rating Scale.[9]

Table: **Clinical features selected for the final model and their relative contributions to the model's precision**

these 11 parameters available in the early stages of the disease. We have also developed a what-if analysis functionality, to explore how feature changes could influence subgroup designation.

Discussion

Researchers and clinicians have long sought a reliable method to identify the subgroups existing within the ALS population. Knowledge of the ALS substructure would improve understanding of the clinical heterogeneity associated with this fatal neurodegenerative disease. By extension, such knowledge would enhance patient care and provide insights into the underlying pathological mechanisms.[22–30] Here, we used a machine-learning approach to identify such subtypes within a large cohort of patients with ALS and replicated our findings in an independent cohort. This data-driven approach confirmed the existence of subtypes within the ALS disease spectrum. Interestingly, these subtypes roughly corresponded to those previously defined by the Chiò classification system,[13] showing the schema's utility. Unlike other subtyping approaches, the Chiò classification system relies on the patient's clinical data collected during the first year of illness.[13] This 1-year observation period allows the disease's symptoms to manifest more clearly and enables the clinician to assess the progression rate more accurately. Although disease progression is a fundamental feature of ALS, it is not typically used in determining the disease subtype.

The primary obstacles to deciphering the clinical heterogeneity observed among patients with ALS have been the absence of a sufficiently large dataset and the inability to analyse multidimensional relationships. To address these issues, we used data from two large, population-based registries that had enrolled patients with ALS over several decades. These registries collected data throughout the patient's illness and, overall, they contained nearly 300 000 pieces of information that we used for our categorisation efforts. Our results highlight the value of disease registries that capture deep phenotypes across an entire catchment area. Previous efforts to catalogue the various subgroups of ALS relied on a small number of clinical features, such as family history or site of symptom onset.[2–5] Although clinically useful, these univariate or bivariate classification systems do not capture the complicated clinical patterns that exist within the ALS population. By contrast, the machine-learning algorithms we applied were adept at deciphering complex and multifaceted relationships. Indeed, the 11 features selected by the supervised model have not been previously combined to predict ALS subtypes.

Our semi-supervised approach, based on a neural network and UMAP, is similar to work published by Sainburg and colleagues. Remarkably, our unsupervised and semi-supervised machine-learning algorithms defined the same subgroups outlined by Chiò and colleagues[13] in their 2011 classification system. This similarity might not be completely surprising in the context of our semi-supervised approach because the same clinical-type-at-1-year patient labels were used to assist the neural network-UMAP clustering. We do not assert that our machine-learning approach is better at identifying categories than experienced ALS neurologists are. Instead, we validated the Chiò classification system using a data-driven approach and provided prima facie evidence that this schema captures the ALS population's substructure. Classification based on other schemes, such as the El Escorial,[5] Milano-Torino,[8] and King's systems,[3] did not help to assign patients to a disease subtype (figure 3).

Nevertheless, our machine-learning algorithm provides opportunities to improve and refine the Chiò classification system, especially as the pyramidal and flail leg ALS subtypes might not be as distinct from each other as other subtypes are. This finding was unexpected, because these patients are easily distinguished from each other in the clinic, highlighting machine-learning's ability to provide new and essential insights into a complex disease, and also offers a novel starting point for exploring the neurobiology underlying the pyramidal and flail leg ALS variants.

Having established that the six subtypes outlined by the Chiò classification system reflected the correct substructures of ALS, we next considered how clinicians and researchers could use this information. The ability to assign patients to subgroups at an early disease stage helps to unravel the disease's clinical heterogeneity and helps in discussions with newly diagnosed individuals about the probable disease course and prognosis. Outcome data from negative clinical trials could be reanalysed for a therapeutic effect limited to one or two subgroups. A similar approach has been successful

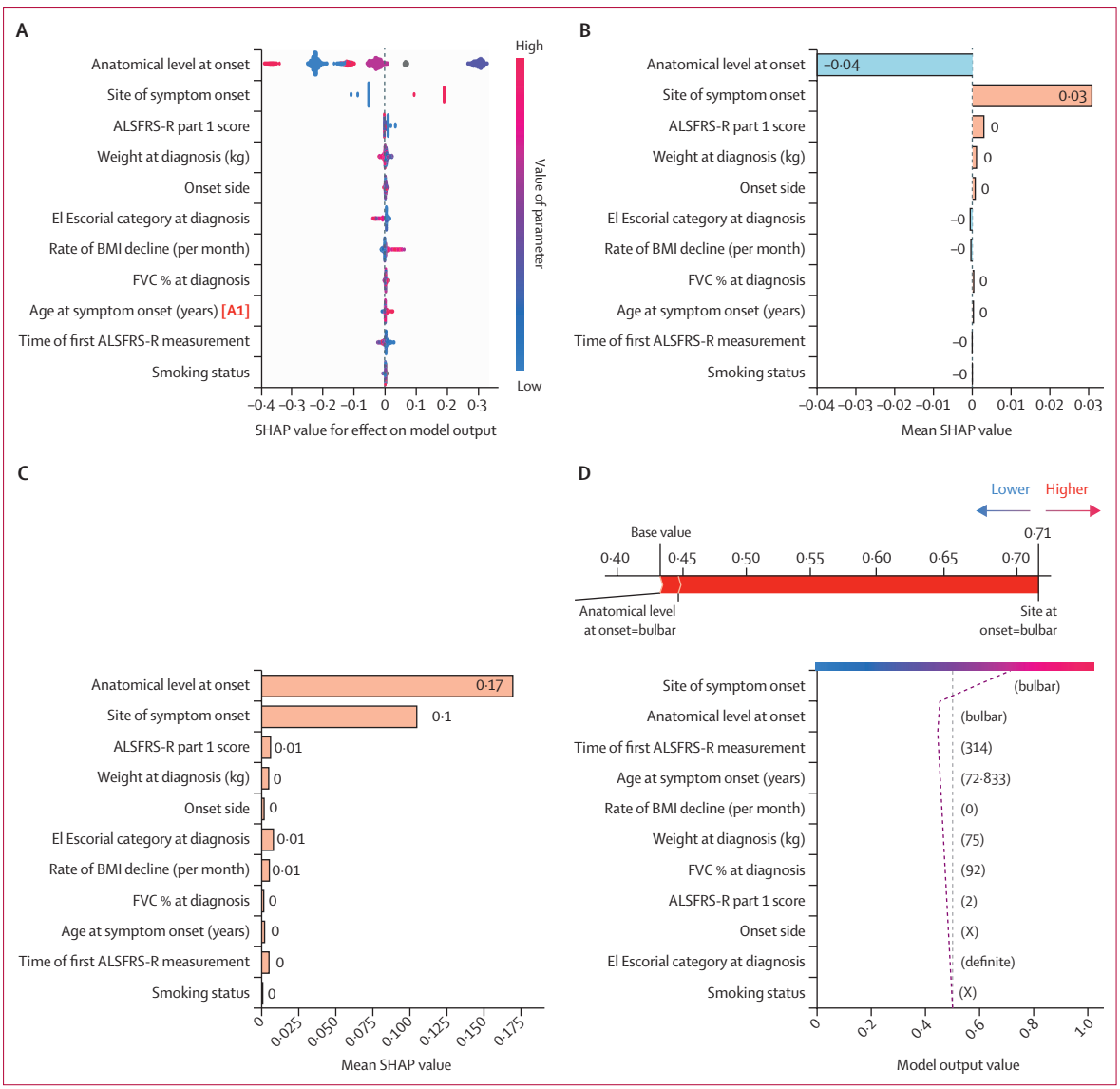

***Figure 5*: The 11 features used in the supervised machine-learning model to predict ALS clinical subtype**
The unit for the time of the first ALSFRS-R measurement was days into illness. (A) Distribution of the 11 features that had the most substantial effect on the predictive value of the classification model over all subtype classes. Each point represents a patient and the amount of effect on model output for each feature depends on its SHAP value. For example, the effect of the rate of BMI decline feature on model output is large when the patient has high values for the rate of BMI decline (in red) as compared to its low values (in blue). The mean of the SHAP values (B) and the mean of the absolute of the SHAP values (C) for the top 11 features, ranked from most important at the top, to least important at the bottom. (D) Force plot (top) and decision plot (bottom) illustrating the influence of each feature on the model's prediction for a single patient with the bulbar subtype of ALS, with unknown onset side and smoking status. The grey dotted line represents the model's base value, whereas the purple dotted line represents the model's prediction and shows how—beginning at the bottom—the SHAP values (ie, feature effects) accumulate from the base value to arrive at the model's final score. The predicted probability that this patient had the bulbar subtype of ALS was 0·71, driven predominantly by the patient's bulbar site of symptom onset, and driven only slightly by their smoking status and El Escorial category at diagnosis (for further examples, see https://share.streamlit.io/anant-dadu/machinelearningforals/main). ALS=amyotrophic lateral sclerosis. ALSFRS-R=Revised ALS Functional Rating Scale.[9] BMI=body-mass index. FVC=forced vital capacity. SHAP=Shapley Additive Explanations.[21]

in Parkinson's disease.[31] Genetic heterogeneity also diminishes our ability to implicate new loci in the disease's pathogenesis using genome-wide association analysis. Including the subgroup as a covariate or restricting the search to a single subtype might resolve this issue by focusing gene-finding efforts within a more homogeneous patient population.

It has not escaped our attention that the topology representation of the ALS subtypes produced by the machine-learning algorithm resembles the CNS. We observed this pattern most clearly in figure 2. The bulbar subtype delineates the cerebrum, and the spinal cord is represented by a long tail running successively from flail arm, pyramidal, classical, to flail leg subtypes. We speculate

that this arrangement hints at a broader anatomical organisation within the ALS spectrum, perhaps reflecting subtle differences of the motor neuron subtypes within each segment of the CNS and differing susceptibilities to pathogenic mechanisms of neurodegeneration.

Our study has several limitations. First, machine-learning algorithms can identify patterns within a dataset even if no such pattern exists. Such overfitting of the model is an inherent problem with this statistical method, and the most legitimate remedy is to attempt replication in an independent dataset. We therefore replicated our findings in an independent, population-based cohort, which yielded remarkably similar outcomes to the discovery cohort, showing the robustness of our approach. Second, the handling of missing data is increasingly recognised as a crucial constraint of machine learning. Our data were remarkably complete, as shown in the exploratory data analysis notebooks. Nonetheless, as with any real-life clinical dataset, information was missing for some parameters, and we aimed to be transparent and cautious in handling these issues. Third, our modelling might have a bias, because we used the same set of patients used by Chiò and colleagues[13] to define their subtypes in their 2011 study. However, it is unlikely that the use of this case series led to sampling bias, because the clinical information used to create the models is standard across the ALS field. Furthermore, population-based registries decrease the possibility of sampling bias because they capture every case within a catchment area. We also replicated our initial findings in an independent cohort that was not used in Chiò and colleagues' 2011 study,[13] confirming that the clusters identified by the data-driven approach did not arise from spurious within-patient associations between variables in the discovery cohort. Nevertheless, both our discovery and replication data originated from the northern Italian population. Additional studies in other countries are required to rule out the possibility of population bias and to test our approach's generalisability. Such data will have to be collected anew, as there is insufficient information to determine the Chiò classification of samples in retrospective data repositories, such as the Pooled Resource Open-Access ALS Clinical Trials Database.[32]

For the **code** see https://github.com/ffaghri1/ALS-ML

For more on **Biowulf** see http://hpc.nih.gov

For the **Pooled Resource Open-Access ALS Clinical Trials Database** see https://ncri1.partners.org/proact

Like other statistical systems, machine-learning algorithms are only practical if they can be applied broadly, and to facilitate this, we have established an interactive website so that physicians can enter a patient's characteristics to predict their ALS subtype. We have made our programming code publicly available so that other researchers can apply it and modify it as our understanding of ALS and machine-learning approaches evolve. Although our current categorisation approach is robust, we anticipate that it will improve over time to the point that it becomes a valuable tool for clinicians helping patients with ALS. Here, we provide an early demonstration of machine learning's ability to unravel highly complex and interrelated disease systems such as ALS.

Contributors

ACh and BJT designed and oversaw the study. FF, FB, MAN, RHC, JM, BJT, and ACh did the primary interpretation of the data. FF and AD designed and implemented the interactive website. FF and BJT wrote the manuscript. ACh and JM made major contributions to manuscript editing. EZ, IM, LM, RV, ACan, CM, ACal, JM, and ACh recruited and phenotyped the study participants. All authors contributed to and critically reviewed the final version of the manuscript. FF, BJT, JM, and ACh verified the data. All authors had access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

BJT holds patents on the clinical testing and therapeutic intervention for the hexanucleotide repeat expansion of *C9orf72* (patent numbers EP2751284A1, CA2846307A, and 20180187262); received research grants from the Myasthenia Gravis Foundation, ALS Association, US Center for Disease Control and Prevention, US Department of Veterans Affairs, MSD, and Cerevel Therapeutics; receives funding through the Intramural Research Program at the US National Institutes of Health (NIH), is on the scientific advisory committee of the American Neurological Association, is an associate editor of *Brain*, and is on the editorial boards of *Journal of Neurology, Neurosurgery, and Psychiatry*, *Neurobiology of Aging*, and *eClinicalMedicine*. JM received research grants from the Fondazione Italiana di Ricerca per la Sclerosi Laterale Amiotrofica, Agenzia Italiana del Farmaco, Italian Ministry of Health, Emilia Romagna Regional Health Authority, and Pfizer. ACh received research funding and honoraria for lectures from Biogen; sits on advisory boards for Mitsubishi Tanabe Pharma, Roche, Denali Therapeutics, Cytokinetics, Biogen, Amylyx Pharmaceuticals, and Sanofi; and participates in data safety monitoring boards for Lilly and AB Science. RV received research scholarship funding from the Rotary Club (global grant GG2094854). FF is employed by Data Tecnica International. MAN is employed by Data Tecnica International and is an adviser for Clover Therapeutics and Neuron23. AD is employed by Data Tecnica International. All other authors declare no competing interests.

Data sharing

Code for preprocessing and prediction is available online. The PARALS and ERRALS registry datasets are not publicly available at the current time, because all research or research-related activities that involve an external party might, at the discretion of the University of Turin or the University Hospital of Modena, require a written research agreement to define obligations and manage risks. To request access to the data, please contact Adriano Chiò (adriano.chio@unito.it) and Jessica Mandrioli (mandrioli.jessica@aou.mo.it). For the PARALS and the ERRALS exploratory data analysis reports, see appendix 3; for further information contact traynorb@mail.nih.gov.

Acknowledgments

We thank staff at the NIH Laboratory of Neurogenetics for their collegial support and technical assistance. This study used the Biowulf Linux cluster high-performance computational capabilities at the NIH. This work was supported by the NIH Intramural Research Program, the US National Institute on Aging (Z01-AG000949–02; funding given to BJT), the Italian Ministry of Health (grant RF-2016–02362405, given to ACh), the European Commission's health Seventh Framework Programme (FP7/2007–2013, under grant agreement 259867 given to ACh), and the Joint Programme–Neurodegenerative Disease Research (funding from the Strength, ALS-Care, and BRAIN-MEND projects given to ACh). This study was funded by a Department of Excellence grant given to ACh by the Italian Ministry of Education, University, and Research, and by the Rita Levi Montalcini Department of Neuroscience, University of Torino, Italy. ERRALS was supported by a grant given to JM by the Emilia Romagna Regional Health Authority. FF, MAN, and AD's participation in this study was part of a competitive contract between Data Tecnica International and NIH.

References

1 Hirtz D, Thurman DJ, Gwinn-Hardy K, Mohamed M, Chaudhuri AR, Zalutsky R. How common are the "common" neurologic disorders? *Neurology* 2007; **68:** 326–37.

2 Byrne S, Bede P, Elamin M, et al. Proposed criteria for familial amyotrophic lateral sclerosis. *Amyotroph Lateral Scler* 2011; **12:** 157–59.

3 Roche JC, Rojas-Garcia R, Scott KM, et al. A proposed staging system for amyotrophic lateral sclerosis. *Brain* 2012; **135:** 847–52.

4 de Carvalho M, Dengler R, Eisen A, et al. Electrodiagnostic criteria for diagnosis of ALS. *Clin Neurophysiol* 2008; **119:** 497–503.

5 Brooks BR. El Escorial World Federation of Neurology criteria for the diagnosis of amyotrophic lateral sclerosis. Subcommittee on Motor Neuron Diseases/Amyotrophic Lateral Sclerosis of the World Federation of Neurology Research Group on Neuromuscular Diseases and the El Escorial "Clinical limits of amyotrophic lateral sclerosis" workshop contributors. *J Neurol Sci* 1994; **124** (suppl)**:** 96–107.

6 Piemonte and Valle d'Aosta Register for Amyotrophic Lateral Sclerosis (PARALS). Incidence of ALS in Italy: evidence for a uniform frequency in Western countries. *Neurology* 2001; **56:** 239–44.

7 Mandrioli J, Biguzzi S, Guidi C, et al. Epidemiology of amyotrophic lateral sclerosis in Emilia Romagna Region (Italy): a population based study. *Amyotroph Lateral Scler Frontotemporal Degener* 2014; **15:** 262–68.

8 Chiò A, Hammond ER, Mora G, Bonito V, Filippini G. Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* 2015; **86:** 38–44.

9 Cedarbaum JM, Stambler N, Malta E, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *J Neurol Sci* 1999; **169:** 13–21.

10 Zheng A, Casari A. Feature engineering for machine learning: principles and techniques for data scientists. Sebastopol, CA: O'Reilly Media, 2018.

11 Han J, Kamber M, Pei J. Data mining: concepts and techniques, 3rd edn. Burlington, MA: Morgan Kaufmann Publishers, 2012.

12 Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak* 2016; **16** (suppl 3)**:** 74.

13 Chiò A, Calvo A, Moglia C, Mazzini L, Mora G. Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study. *J Neurol Neurosurg Psychiatry* 2011; **82:** 740–46.

14 McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw* 2018; **3:** 861.

15 Sainburg T, McInnes L, Gentner TQ. Parametric UMAP embeddings for representation and semi-supervised learning. *Neural Comput* 2021; **33:** 2881–907.

16 Rokach L. Ensemble-based classifiers. *Artif Intell Rev* 2010; **33:** 1–39.

17 Makarious MB, Leonard HL, Vitale D, et al. GenoML: automated machine learning for genomics. *arXiv* 2021; published online March 4. https://arxiv.org/abs/2103.03221v1 (preprint).

18 Breiman L. Random forests. *Mach Learn* 2001; **45:** 5–32.

19 Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. 31st International Conference on Neural Information Processing Systems 2017; Dec 4–9, 2017.

20 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining Conference on Knowledge Discovery and Data Mining. Aug 13–17, 2016.

21 Lundberg SM, Lee S. A unified approach to interpreting model predictions. 31st International Conference on Neural Information Processing Systems 2017; Dec 4–9, 2017.

22 Kueffner R, Zach N, Bronfeld M, et al. Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Sci Rep* 2019; **9:** 690.

23 Küffner R, Zach N, Norel R, et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat Biotechnol* 2015; **33:** 51–57.

24 Tang M, Gao C, Goutman SA, et al. Model-based and model-free techniques for amyotrophic lateral sclerosis diagnostic prediction and patient clustering. *Neuroinformatics* 2019; **17:** 407–21.

25 Grollemund V, Chat GL, Secchi-Buhour MS, et al. Development and validation of a 1-year survival prognosis estimation model for amyotrophic lateral sclerosis using manifold learning algorithm UMAP. *Sci Rep* 2020; **10:** 13378.

26 Beaulieu-Jones BK, Greene CS, Pooled Resource Open-Access ALS Clinical Trials Consortium. Semi-supervised learning of the electronic health record for phenotype stratification. *J Biomed Inform* 2016; **64:** 168–78.

27 Elamin M, Bede P, Montuschi A, Pender N, Chio A, Hardiman O. Predicting prognosis in amyotrophic lateral sclerosis: a simple algorithm. *J Neurol* 2015; **262:** 1447–54.

28 Ong ML, Tan PF, Holbrook JD. Predicting functional decline and survival in amyotrophic lateral sclerosis. *PLoS One* 2017; **12:** e0174925.

29 Pfohl SR, Kim RB, Coan GS, Mitchell CS. Unraveling the complexity of amyotrophic lateral sclerosis survival prediction. *Front Neuroinform* 2018; **12:** 36.

30 Westeneng HJ, Debray TPA, Visser AE, et al. Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *Lancet Neurol* 2018; **17:** 423–33.

31 Leonard H, Blauwendraat C, Krohn L, et al. Genetic variability and potential effects on clinical trial outcomes: perspectives in Parkinson's disease. *J Med Genet* 2020; **57:** 331–38.

32 Atassi N, Berry J, Shui A, et al. The PRO-ACT database: design, initial analyses, and predictive features. *Neurology* 2014; **83:** 1719–25.