Predicting Clinical Outcomes of Amyotrophic Lateral Sclerosis Progression via Logistic Regression and Deep-Learning Multilayer Perceptron Approaches Using Routine Visits Data

Alessandro Guazzo[1,*], Michele Atzeni[1,*], Elena Idi[1,*], Isotta Trescato[1], Erica Tavazzi[1], Enrico Longato[1], Umberto Manera[2], Adriano Chiò[2], Marta Gromicho[3], Inês Alves[3], Mamede de Carvalho[3], Martina Vettoretti[1], and Barbara Di Camillo[1,4]

[1] Department of Information Engineering, University of Padova, Padua, Italy.
[2] Department of Neurosciences “Rita Levi Montalcini”, University of Turin, Turin, Italy.
[3] Faculdade de Medicina, Instituto de Medicina Molecular João Lobo Antunes, Universidade de Lisboa, Lisbon, Portugal.
[4] Department of Comparative Biomedicine and Food Science, University of Padova, Padua, Italy.

*These authors contributed equally

Keywords: *Amyotrophic Lateral Sclerosis, Multi Layer Perceptron, Multinomial Logistic Regression*

Abstract.

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease that results in death within a short time span (3-5 years). One of the major challenges in treating ALS is its highly heterogeneous disease progression and the lack of effective prognostic tools to forecast it. The main aim of this study was, then, to test the feasibility of predicting relevant clinical outcomes that characterize the progression of ALS with a two-year prediction horizon via artificial intelligence techniques using routine visits data. Three classification problems were considered: predicting death (binary problem), predicting percutaneous endoscopic gastrostomy (PEG) or death (multiclass problem), and predicting non-invasive ventilation (NIV) or death (multiclass problem). Two supervised learning models, a multinomial logistic regression (MLR) and a deep learning multilayer perceptron (MLP), were trained ensuring technical robustness and reproducibility. Results showed that predicting death was the most successful task, with both models yielding F1-scores and accuracy above 0.7. However, predicting PEG or NIV alongside death proved to be much more challenging (F1-scores and accuracy in the 0.6-0.7 interval) In conclusion, predicting clinical outcomes of ALS progression proved to be feasible. However, a complex learning technique, such as the MLP, led to comparable results with respect to the simpler MLR. Such invariance to model capacity suggests that a cap on how much information can be extracted from the database might have been reached, e.g., due to the intrinsic difficulty of the prediction tasks at hand, or to the absence of crucial predictors that are, however, not currently collected during routine practice.

1 Introduction

Amyotrophic lateral sclerosis (ALS) is a rapidly progressive neurodegenerative disease that affects motor neurons in the brain and spinal cord, leading to muscle weakness, atrophy, and, ultimately, paralysis. The disease typically results in death within a relatively short span of 3-5 years, although survival times can vary widely and depend on many factors including age, site of onset, rate of disease progression, and presence of comorbidities [1]. ALS is also characterized by a significant heterogeneity in its progression across the patient population, with some individuals showing a slow progression and others experiencing a rapid decline [2].

One of the major challenges in treating ALS is the lack of effective prognostic tools for predicting disease progression. While there are some clinical factors that have been identified as potential predictors, such as age, site of onset, and functional impairment, these are not always reliable indicators and cannot fully capture the complexity of the disease. Accurate prognostic tools would enable better drug development using cheaper and more accurate clinical trials, as well as providing valuable insights into disease mechanisms and manifestations.

In recent years, artificial intelligence (AI) approaches have shown promise for predicting clinical outcomes in various disease contexts, including ALS [3, 4, 5, 6]. By leveraging complex algorithms and large amounts of data, AI-based models can identify patterns and relationships that may not be immediately apparent to human observers. In the case of ALS, this could potentially lead to more accurate prognostic tools that could help clinicians tailor treatment plans to individual patients.

The objective of this research work is to use AI-based approaches to predict relevant clinical outcomes that characterize the progression of ALS. Two types of AI models were developed and tested to predict three clinical outcomes of great interest to clinicians: (i) death within two years from the first visit, (ii) percutaneous endoscopic gastrostomy (PEG) or death (whichever occurs first) within two years from the first visit, and (iii) non-invasive ventilation (NIV) or death (whichever occurs first) within two years from the first visit. Results suggest that while the considered AI approaches had good predictive performance for the binary classification task of predicting death within two years, combining the prediction of death with either NIV or PEG in a multiclass classification problem is more challenging considering the available data. Our study provides valuable insights into the potential and limitations of AI-based prognostic tools for ALS using as input features variables obtained from routine ALS visits.

2 Data and Methods

2.1 *Dataset and preprocessing*

The dataset used in this study was provided by the European Horizon 2020 project BRAINTEASER [7]. The BRAINTEASER project aims to use AI to gain a better understanding of ALS, predict disease progression, and propose interventions to delay its advancement. This involves developing models that can identify and forecast disease outcomes over time for different patient groups, providing support for patient care and clinical trials. Detecting complications during the disease progression is crucial for ALS patients and healthcare professionals.

The ALS dataset, provided within the BRAINTEASER project, includes data coming from two data registries, one Italian and one Portuguese. On one hand, the Italian ALS data registry is based on the Piemonte and Valle d'Aosta Register for Amyotrophic Lateral Sclerosis (PARALS). PARALS is an epidemiologic prospective register that covers two Italian regions (population of 4,476,931 inhabitants according to the 2011 census). Demographic and clinical data from 3,257 ALS patients collected from January 1, 1995, through December 31, 2018 were considered from this registry. On the other hand, the Lisbon ALS registry contains demographic and clinical data from 1,562 ALS patients regularly followed at the ALS clinic at Hospital de Santa Maria, Lisbon (CHULN) since 1995 and last updated in October 2021. The two registries were harmonized to obtain a set of common variables to be used as model inputs. The list of input variables is reported in Table 1. Input data were processed by following the procedure described in [6]. Outcome variables were processed to consider the problem from a classification perspective as an alternative to the survival analysis perspective already explored in [6, 8, 9]. Consequently, the

Table 1: List of variables considered as inputs for the models. Baseline variables were obtained from raw data collected at the first visit meanwhile follow-up variables were obtained from raw data collected at multiple visits within 6 months after the index date.

Section	Sub-section	Variables
Baseline	Demographics	age, sex, instruction level
	Anthropometrics	BMI, weight slope
	ALS Onset and diagnosis	onset type, time since onset, diagnostic delay, prevalent motor neuron involvement
	ALSFRS subscores	breathing, bulbar, lower limbs, trunk, upper limbs
	Genetic mutations	C9orf72, FUS, TARDBP, SOD1
	Previous pathologies	autoimmune disease, cardiac disease, diabetes, dyslipidemia, hypertension, primary neoplasm, stroke, thyroid disorder, surgery before onset, surgery after onset, trauma before onset, trauma after onset, ALS familiar history
	Lifestyle	smoking
Follow-up	ALSFRS progression slopes	breathing, bulbar, lowerlimbs, trunk, upper limbs
	Tests	forced vital capacity

overall dataset was used to derive three sub-datasets, one for each outcome of interest (death, PEG or death, and NIV or death). For the first sub-dataset (N = 2114 subjects), a binary outcome was considered, specifically, the label 1 was assigned to patients that died within two years after the first visit (1059 out of 2114, 50% meanwhile, the label 0 was assigned to those that survived (1055 out of 2114, 50%). The second sub-dataset (N = 2027 subjects) considered two possible events, namely PEG or death, occurring always within two years from the first visit. In this case, the label 1 was assigned to patients for which the first recorded outcome was death (684 out of 2027, 33.7%), the label 2 was assigned to patients for which the first recorded outcome was PEG (455 out of 2027, 22.4%), and the label 0 was assigned to patients that experienced neither of the two events within the two years (888 out of 2027, 43.8%). Finally, also the third sub-dataset (N = 1742 subjects) considered two possible events, namely NIV or death, occurring within two years from the first visit. In this case, the label 1 was assigned to patients for which the first recorded outcome was death (497 out of 1742, 28.5%), the label 2 was assigned to patients for which the first recorded outcome was NIV (592 out of 1742, 34%), and the label 0 was assigned to patients that experienced neither of the two events within the two years (653 out of 1742, 37.5%).

Each of the three sub-datasets was then divided into three subsets: training, validation, and test sets, with each set comprising 70%, 15%, and 15% of the total dataset, respectively.

2.2 *AI model development and evaluation*

Two supervised machine learning approaches were considered to solve the three classification problems, a simple linear approach, namely a Multinomial Logistic Regression with L2 regularization (MLR), and a more complex non-linear approach, namely a Deep Learning Multilayer Perceptron (MLP). To ensure the technical robustness and reproducibility of the study a two-step optimization framework was implemented.

The first step performs hyperparameter optimization by the random search approach for both MLR and MLP. The MLR requires optimization of the regularization parameter C defining the L2 penalty term (uniformly sampled in the range [0.0001, 1.0] according to a logarithmic scale). The MLP requires optimization of various hyperparameters, such as learning rate (sampled from a descending sequence ranging in [0.01, 0.0001], with each number being one-tenth of the previous number), initializer (sampled from all the possible initializers, i.e., random normal, random uniform, glorot normal, he normal, he uniform, variance scaling, orthogonal), activation function (sampled from all the possible activation functions, i.e., relu, tanh, selu, elu), and architecture complexity (funnel-like structure where the number of nodes in the next layer is half the number of nodes in the previous layer, ranging in a maximum starter nodes in [512, 8]). For each random search step, a five-fold cross-validation was performed on the training set to obtain the mean cross-validation cross-entropy loss. The best hyperparameters set was chosen based on the minimum cross-validation cross-entropy loss obtained over 500 random iterations.

The second optimization step consisted of the training, on the whole training set, of several models using the optimal set of hyperparameters (obtained from the previous step) and different random initializations. The best model was chosen by minimizing the cross-entropy loss computed on the validation set among 100 random iterations.

Finally, the optimal model was tested on an independent portion of the data considering as performance evaluation metrics the Area under the Precision-Recall Curve (AUPRC) and the Area Under the Receiver-Operating Curve (AUROC) as well as Precision, Recall, F1-Score, and Accuracy computed after thresholding. For binary classification, the threshold was set using the validation set by testing all the predicted probabilities and selecting the one that maximized the geometric mean between specificity and sensitivity. In the multiclass case, instead, the class was assigned by selecting the one with highest predicted probability in a 1 vs. all fashion.

3 Results and Discussion

The performance metrics obtained with the two considered approaches (MLP and MLR) for the three considered outcomes of interest (Death, PEG or Death, and NIV or Death) are reported and discussed in the following subsections.

3.1 *Death prediction*

The performance metrics obtained when considering the death outcome are reported in Table 2. Both models (MLP and logistic regression) performed well on the independent test set reaching F1-scores well above 0.7. AUPRC and AUROC were good as well (~ 0.8). Overall, the two methods led to comparable results with the MLP achieving slightly higher precision and the LR reaching higher recall instead.

Table 2: Table containing the considered performance metrics (AUPRC, AUROC, Precision, Recall, F1-Score Accuracy) for the ”Death” outcome.

Model	**AUPRC**	**AUROC**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
MLP	0.79	0.81	0.72	0.75	0.74	0.73
LR	0.8	0.81	0.71	0.8	0.75	0.74

3.2 *Death or PEG prediction*

AUPRC, AUROC, precision, recall, and F1-score metrics for predicting Death or PEG are shown in (Table 3). Both the MLP and MLR models show promising predictive performance in the one-versus-all case for predicting the absence of an event versus the occurrence of death and PEG, reaching AUPRC and AUROC of ~ 0.85 and F1-score above 0.7. On the contrary, predicting ”Death” vs. ”PEG” & ”No event” and ”PEG” vs. ”No event” & ”Death” led to general lower predictive performance (AUPRC and F1-Score ~ 0.6).

Table 3: Table containing the performance metrics (AUPRC, AUROC, Precision, Recall, F1-Score and Overall Accuracy) results for the ”No event” vs ”PEG” vs ”Death” classification task.

Model	Outcome	AUPRC	AUROC	Precision	Recall	F1-Score	Accuracy
MLP	No event	0.86	0.85	0.71	0.78	0.75	-
	Death	0.64	0.81	0.63	0.66	0.64	-
	PEG	0.67	0.85	0.73	0.5	0.59	-
							0.68
MLR	No event	0.85	0.85	0.67	0.81	0.73	-
	Death	0.62	0.79	0.60	0.54	0.57	-
	PEG	0.64	0.83	0.72	0.53	0.61	-
							0.66

3.3 *Death or NIV prediction*

Similarly to the other considered outcomes MLP and MLR were assessed using recall, precision, F1-score, and accuracy as evaluation metrics which are reported in Table 4. MLP achieves better performance when predicting the absence of adverse events and instead struggles when distinguishing if NIV or death occurs first. Interestingly, the prediction of NIV is a more challenging task whose difficulties can be related to the more variability in the timing of this intervention with respect to PEG. The MLR, instead, leads to slightly better precision and better results in terms of recall with respect to the MLP. Even if a small improvement has been found when using MLR instead of MLP, the discrimination ability in the prediction of NIV is still low.

Table 4: Table containing the performance metrics (AUPRC, AUROC, Precision, Recall, F1-Score and Overall Accuracy results for the ”No event” vs ”NIV” vs ”Death” classification task.

Model	Outcome	AUPRC	AUROC	Precision	Recall	F1-Score	Accuracy
MLP	No event	0.67	0.78	0.62	0.60	0.61	-
	Death	0.66	0.79	0.58	0.59	0.59	-
	NIV	0.58	0.71	0.56	0.56	0.56	-
							0.59
MLR	No event	0.81	0.86	0.65	0.80	0.72	-
	Death	0.66	0.83	0.62	0.58	0.60	-
	NIV	0.63	0.78	0.61	0.48	0.54	-
							0.63

4 Conclusion

In this work, the data collected by two ALS registries (the Italian PARALS registry and the Lisbon ALS registry) were used to develop prognostic models of relevant clinical outcomes of ALS. Specifically, three outcomes of increasing complexity were considered, namely death (binary outcome), PEG or death (multiclass outcome), and NIV or death (multiclass outcome). Two different modeling approaches were considered: a simple linear approach, i.e., MLR, and a more complex non-linear approach, i.e., MLP. Results suggest that the prediction of such ALS outcomes with a prediction horizon of 2 years after the first visit is feasible. The developed models showed good predictive performance, especially in the binary classification task of predicting death, while predicting multiclass outcomes such as NIV or PEG alongside death proved to be more challenging with the available variables. However, the use of more complex algorithms, such as the MLP, did not provide a significant improvement in performance when compared to linear techniques such as the MLR, and, in the most challenging task of predicting a non-invasive procedure, such as NIV, the simpler MLR approach led to better performance than MLP.

As model performance was comparable between the linear and the non-linear techniques, the main driver of predictive performance might be the information that can be extracted from the available data, which include only variables recorded during routine visits performed at ALS centers. Thereby, other, more outcome-specific information (i.e., more invasive measurements, such as blood or cerebrospinal fluid tests) could be helpful to improve the prediction of ALS outcomes of interest.

In conclusion, our study highlights the potential of AI approaches in predicting relevant clinical outcomes of ALS, particularly for predicting clinically complex events such as death. In the future, further studies may focus on the collection or extraction of time-varying and outcome-specific variables as well as the development of more sophisticated methodologies able to better consider temporal information with the aim of improving the predictive performance of AI-based approaches as well as the implementation of variable ranking techniques to better understand feature importance with respect to the model output.

Conflict of interests

The authors declare that do not have a conflict of interest.

Data Availability

The BRAINTEASER ALS ontology describing the data used for this study is available at the URL: https://zenodo.org/record/7886998#.ZF5s5HZByUl

References

[1] Brown and Al-Chalabi, “Amyotrophic lateral sclerosis,” *N Engl J Med*, vol. 2;377(9):162-172., 2017.

[2] N. Atassi, J. Berry, A. Shui, N. Zach, A. Sherman, E. Sinani, and et al., “The pro-act database: design, initial analyses, and predictive features,” *Neurology*, vol. 83(3):171-8, 2014.

[3] E. Tavazzi, S. Daberdaku, A. Zandonà, R. Vasta, B. Nefussy, C. Lunetta, and et al., “Predicting functional impairment trajectories in amyotrophiclateral sclerosis: a probabilistic, multifactorial model of disease progression,” *Neurology*, vol. 1–21, 2022.

[4] A. S. Martins, M. Gromicho, S. Pinto, M. de Carvalho, and S. C. Madeira, “Learning prognosticmodels using disease progression patterns: Predicting the need for non-invasive ventilation in amyotrophic lateral sclerosis,” *IEEE/ACM Transactions on Computational Biology andBioinformatics*, 2021.

[5] J. Ackrivo, J. Hansen-Flaschen, E. P. Wileyto, R. J. Schwab, L. Elman, and S. M. Kawut, “Development of a prognostic model of respiratory insufficiency or death in amyotrophic lateral sclerosis,” *European Respiratory Journal*, vol. 53, 2019.

[6] I. Trescato, A. Guazzo, E. Longato, E. Hazizaj, C. Roversi, E. Tavazzi, and et al., “Baseline machine learning approaches to predict amyotrophic lateral sclerosis disease progression notebook for the idpp lab on intelligent disease progression prediction at clef 2022,” 2022.

[7] Feb 2023. [Online]. Available: https://brainteaser.health/

[8] A. Guazzo, I. Trescato, E. Longato, E. Hazizaj, D. Dosso, G. Faggioli, and et al., “Intelligent disease progression prediction: Overview of idpp@clef 2022,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer International Publishing, 2022, pp. 395–422.

[9] ——, “Overview of idpp@ clef 2022: the intelligent disease progression prediction challenge,” in *CLEF*, 2022, pp. 1613–0073.