Journal of High School Science

Causal impacts of blood biomarkers on Amyotrophic Lateral Sclerosis progression: a novel application of double Machine Learning

Yilmaz B

Submitted: September 14, 2024, Revised: version 1, November 6, 2024
Accepted: November 7, 2024

Abstract
Amyotrophic Lateral Sclerosis (ALS) is a rapidly progressing neurodegenerative disease with no known cure and limited treatment options. While previous research has identified various blood biomarkers correlated with ALS progression, a deeper understanding of how these biomarkers influence disease progression is essential for developing targeted and effective therapeutic strategies. This study is the first to apply Double Machine Learning (DML) to blood biomarker data in ALS research, enabling robust statistical inference on treatment effects within high-dimensional clinical data. Using data from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database, key biomarkers were identified with significant treatment effects on ALS progression. Protective effects were observed from albumin (treatment effect = -0.18, $p = 0.001$), creatinine (treatment effect = -0.05, $p < 0.001$), and lymphocytes (treatment effect = -0.07, $p = 0.004$). In contrast, accelerating effects were linked to elevated calcium (treatment effect = 8.13, $p < 0.001$), bicarbonate (treatment effect = 0.60, $p < 0.001$), and phosphorus (treatment effect = 6.06, $p < 0.001$). Additional biomarkers such as glucose (treatment effect = 0.27, $p = 0.016$), chloride (treatment effect = -0.39, $p < 0.001$), and creatine kinase (treatment effect = -0.01, $p = 0.008$) were also identified as significant factors influencing progression. These findings underscore the importance of distinguishing between correlation and statistical causality, providing valuable insights for designing targeted interventions in ALS treatment. Additionally, clinical factors such as the initial Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) score, bulbar-related symptoms, recent symptom onset, gastrointestinal complications, and time since the first symptoms were identified were critical determinants of disease progression. This application of DML in ALS research enhances our understanding of influential drivers in disease progression and lays the groundwork for future clinical studies and therapeutic innovations.

Keywords
Machine Learning, ALS, Biomarkers, Predictive modeling, Neurodegenerative disease, ALS Functional Rating Scale (ALSFRS), Causal inference, Treatment effect, Double Machine Learning (DML), High-dimensional data

Berke Yilmaz, Tesla STEM High School, 4301 228th Ave NE, Redmond, WA 98053, USA.
berkeyilmaz.us@gmail.com

1. Introduction

1.1 Background

Amyotrophic Lateral Sclerosis (ALS), also known as Lou Gehrig's disease, is a fatal neuromuscular disorder characterized by the degeneration of both upper and lower motor neurons. Most patients survive only two to five years after diagnosis (1). While genetic factors account for 5-10% of cases, the majority are sporadic with no identifiable cause. Approximately 5,000 new cases are diagnosed in the U.S. each year (2). According to Mehta et al. (3), data from the National ALS Registry estimates that there are ~ 24,821 ALS cases in the U.S., with an upper estimate of 31,843, corresponding to a prevalence rate of 7.7 to 9.9 per 100,000 residents.

Arthur et al. (4) projected a significant global rise in ALS cases, estimating an increase from 222,801 in 2015 to 376,674 by 2040, representing a 69% growth in the number of individuals affected. ALS also places a significant financial burden on patients. Achtert and Kerkemeyer (5) reviewed various studies focused on the costs per ALS patient, reporting a range from €9,741 to €114,605 (~ $10,715 to $126,065). In the U.S., the average annual cost for ALS patients is ~ $70,000 (2).

There is currently no cure for ALS. The primary focus of treatment is to slow disease progression and enhance the quality of life for patients by providing the necessary equipment to help them maintain as much normalcy as possible. Riluzole, Edaravone, and Sodium Phenylbutyrate/Taurursodiol are the only drugs approved for ALS treatment, offering moderate effects on slowing disease progression (6).

Given the significant impact of ALS and the limitations of current treatments, understanding the factors that influence disease progression is critical for advancing clinical care. While environmental, genetic, and blood biomarkers are all implicated in ALS, blood biomarkers offer unique insights for developing targeted therapies and improving early diagnosis. Identifying biomarkers with statistically inferred causal associations allows researchers to pinpoint key drivers of progression, facilitating more precise interventions and improved patient management. Additionally, tracking these biomarkers over time can yield valuable insights into treatment efficacy and support personalized care. This study examines the inferred impact of blood biomarkers and clinical factors on ALS progression.

In recent years, machine learning-based causal inference techniques have become increasingly prominent in healthcare research, especially for analyzing complex, high-dimensional datasets. Studies by Sanchez et al. (7) and Feuerriegel et al. (8) highlight the importance of methods like Double Machine Learning (DML) for producing robust insights that support clinical decision-making. Given that ALS progression is influenced by multiple, interrelated factors, causal inference methods are essential for identifying biomarkers that may influence disease trajectory.

Among various causal inference methods, DML has emerged as particularly effective for adjusting for confounding factors and isolating treatment effects. This approach is especially suited to high-dimensional datasets, where relationships between variables are complex. DML's versatility has been demonstrated across healthcare applications, enabling

detailed analysis of data while addressing confounding influences. For example, DML has been used to evaluate the Women, Infants, and Children (WIC) program's effects on infant health outcomes, helping researchers identify variations based on risk profiles (9). In oncology, DML has advanced pharmacogenomic research by assessing epithelial-mesenchymal transition (EMT) signatures in cancer cell lines, revealing specific drug susceptibilities and informing precision treatment strategies (10).

These examples underscore DML's flexibility in healthcare settings, where accurate estimation of treatment effects is critical. By applying DML to blood biomarker data, this study similarly aims to deepen the understanding of ALS progression, offering a statistically rigorous foundation for targeted interventions.

In this manuscript, *causality* refers to the quantitative predictive value of a blood biomarker on the progression of ALS. This concept of causality allows us to analyze how the system might respond to interventions, providing actionable insights into potential treatment effects by predicting outcomes such as efficacy and toxicity (8). This approach aligns with the framework of *statistical* causality, which emphasizes estimating causal effects and predicting treatment outcomes in a systematic manner (11).

However, this statistical approach to causality does not—and cannot—fully differentiate between correlation and causation in the molecular biology context, where causation implies an understanding of the mechanism of action (12). In molecular biology, causation typically involves a signaling pathway where molecular or chemical changes lead to a pharmacological effect. The complexity of these pathways, along with phenomena such as pleiotropy—where a single factor impacts multiple physiological processes—and systemic redundancies, means that causation often involves a network of related events. This complex interplay helps ensure system robustness but also makes it challenging to pinpoint a single cause for an observed effect. In such contexts, causative events may be necessary or sufficient to produce an outcome, with the progression from cause to effect being grounded in well-defined molecular pathways.

In contrast, the statistical approach employed in this study leverages DML to estimate treatment effects and provide insights into the relationships between blood biomarkers and ALS progression. While not aimed at uncovering molecular mechanisms, this statistical framework allows us to identify potential causal associations by adjusting for confounding factors, thereby isolating effects that could inform targeted interventions.

Recent literature, including the Evidence-Based Medicine+ approach, underscores the importance of combining clinical research with mechanistic evidence, as articulated by the Russo and Williamson Thesis (13). Integrating these perspectives provides a more comprehensive understanding of causality, moving beyond traditional evidence-based methods to incorporate the explanatory insights offered by mechanisms (14). This study, however, focuses on evidence-based methods within the DML framework to infer statistical causality.

1.2 Pathophysiology of ALS

ALS is a neurodegenerative disease characterized by the progressive degeneration of motor neurons, leading to a loss of muscle control and eventual paralysis. Figure 1 provides a visual comparison between a normal motor neuron and an ALS-affected motor neuron, illustrating the stark differences in structure and the impact on associated muscle tissue.

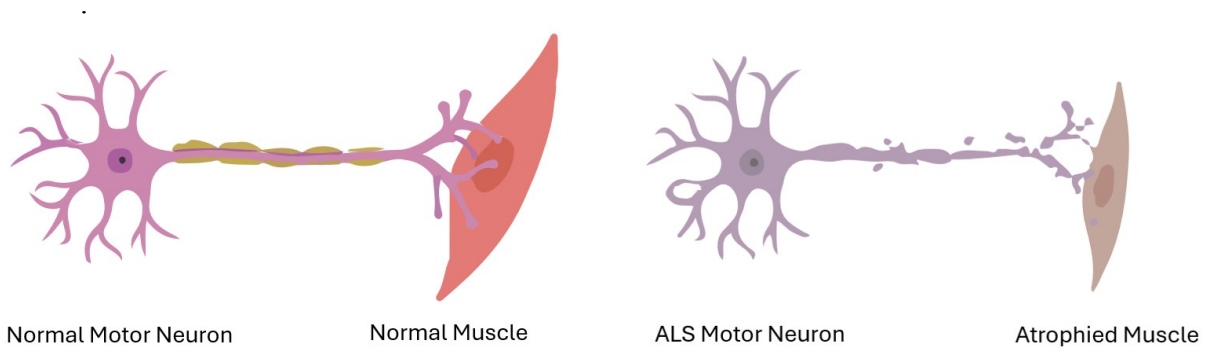

Figure 1. Comparison of a normal motor neuron displayed on the left and an ALS motor neuron displayed on the right.

The healthy motor neuron, as shown on the left in Figure 1, maintains a strong connection to a normal muscle. The neuron's dendrites and axon are well-defined, representing an intact communication pathway between the nervous system and the muscle, which supports normal muscle function and strength. In contrast, the image of a motor neuron affected by ALS, shown on the right in Figure 1, exhibits significant structural degradation, characteristic of the disease. The accompanying muscle is depicted in an atrophied state, which appears weakened and diminished, reflecting the loss of muscle mass and function as the motor neuron deteriorates and loses its ability to properly stimulate the muscle.

1.3 Research objectives

The primary objective of this research was to estimate how different interventions, as suggested by key biomarkers, may influence the progression of sporadic ALS. By identifying these biomarkers, the study aimed to inform strategies that could potentially slow disease progression and improve life expectancy, offering patients and caregivers enhanced options for disease management.

Determining which biomarkers are linked to accelerated or slowed ALS progression is crucial for developing personalized and effective interventions. This research focused on isolating biomarkers that indicate actionable pathways, providing a foundation for targeted therapeutic approaches.

By applying advanced statistical methods, this study sought to understand how potential interventions could alter the trajectory of ALS progression, thereby contributing to more precise treatment strategies and improved patient outcomes.

2. Methods

Figure 2 provides a visual overview of the methodology employed in this study, which leverages the Double Machine Learning (DML) framework for statistical causal inference in ALS progression (15). The process begins with raw data collection and culminates in the extraction of actionable insights through the interpretation of model results. Each step in the methodology is designed to ensure a thorough and precise analysis, ultimately aiming to uncover the relationships between biomarkers and ALS progression.

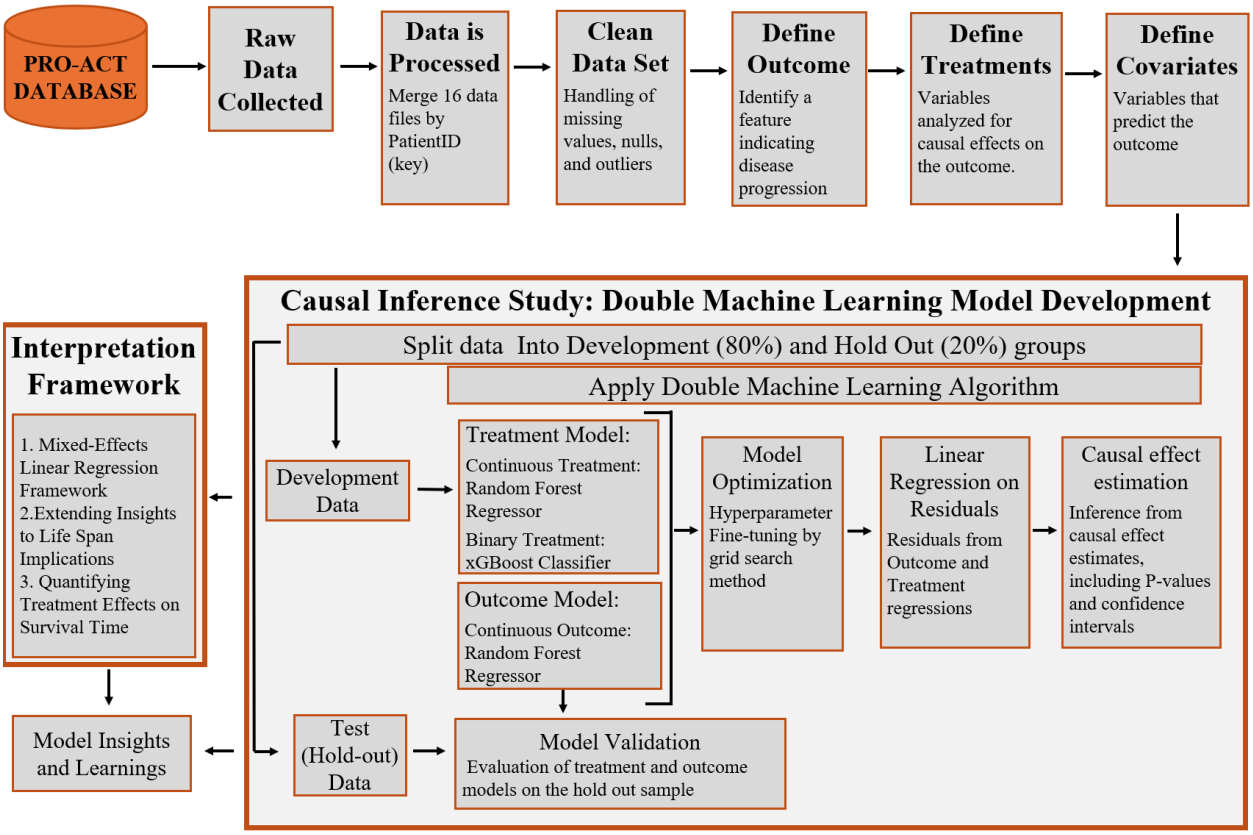

Figure 2. Overview of the methodology for analyzing average treatment effects of biomarkers on ALS progression using advanced machine learning methods. The process includes data collection and processing, defining the outcome and confounders, and applying the Double Machine Learning (DML) algorithm. The framework also involves model validation, insights interpretation, and quantification of treatment effects on survival time.

2.1 Data collection and preprocessing

In the early phase of the study, an extensive effort was undertaken to obtain and integrate metadata from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database (16). Established in 2012, PRO-ACT is renowned for being the most extensive source of anonymized ALS patient data, encompassing ~ 11,600 patient records and ~ 10 million data points derived from 29 clinical trials. The available data tables include Adverse Events, Forced Vital Capacity (FVC),

Slow Vital Capacity (SVC), Laboratory test results of various biomarkers, ALS History, Vital Signs, Death Data, Demographics, ALS FRS, El Escorial Criteria, Family History, Hand Grip Strength, Muscle Strength, Riluzole, and Concomitant Medications. These tables were merged using Patient ID as the key, resulting in a comprehensive dataset with a wide array of patient information.

The data was then processed to identify, correct and/or impute missing values, nulls, and outliers, ensuring a clean dataset ready for analysis. After extensive data cleaning, the dataset consisted of 249 features and 5,865 patients.

2.2 Defining the outcome variable

Once the data was cleaned, the next step was to define the outcome (dependent) variable, which in this study represented disease progression. For this purpose, the ALS Functional Rating Scale (ALSFRS), a comprehensive measure of functional decline in ALS patients, was selected. The ALSFRS is a clinical tool used by healthcare providers to monitor the decline in functional abilities of ALS patients over time. It evaluates various aspects of a patient's daily functioning, including speech, salivation, swallowing, handwriting, utensil use, dressing, hygiene, bed mobility, walking, stair climbing, and breathing. The scale consists of 12 items, each scored from 0 to 4, with higher scores reflecting better functional capacity. A maximum score of 48 indicates normal function, while lower scores correspond to increasing levels of disability.

The ALSFRS has been validated for its reliability and sensitivity to changes in motor and pulmonary function, making it a robust benchmark for evaluating disease progression in ALS ("The Amyotrophic Lateral Sclerosis Functional Rating Scale," 1996). A nominal decrease in ALSFRS provides a consistent benchmark for evaluating the estimated treatment effects, establishing a reliable foundation for analysis within the Double Machine Learning (DML) framework.

2.3 Data analysis

The data analysis process was carefully designed to leverage the Double Machine Learning (DML) framework, with the goal of estimating and isolating treatment effects on ALS progression. This approach enables an assessment of how ALS progression may respond to potential interventions.

2.3.1 *Feature engineering*

The initial phase began with feature engineering, expanding the Pro-ACT dataset from 147 columns to 415 columns by constructing additional variables relevant to ALS progression. These newly constructed features captured the time series aspects of the biomarkers, such as measurements at different time points and changes in values over specific periods.

2.3.2 *Data preprocessing*

Following feature engineering, univariate analysis was performed to explore each feature independently, examining distributions, central tendencies, and variability. This step provided crucial insights into the data, enabling the identification of outliers and ensuring the quality of features for subsequent analysis. Data cleaning activities were employed such as removing features with high percentages of missing values, imputing missing data, and processing outliers. These processes ensured

the dataset was of high quality and ready for subsequent analysis. In this step, features with more than 30% missing values were removed, resulting in a feature list of 249 columns.

2.3.3 *Feature selection and dimensionality reduction*

There are nuanced perspectives on the need for dimensionality reduction in Double Machine Learning (DML). Chernozhukov et al. (15), in the original paper introducing DML, highlight its capability to handle high-dimensional data effectively without requiring dimensionality reduction. Nonetheless, some studies suggest that reducing the feature set can improve model performance and facilitate interpretation, particularly in cases with complex, high-dimensional datasets where noise or redundancy might affect results. Hence, carefully selecting variables and tuning algorithms can significantly enhance the accuracy and robustness of treatment effect estimation (17-19). Therefore, in this analysis, feature selection was performed progressively, using Principal Component Analysis (PCA), bivariate analysis, stepwise regression, and a Random Forest algorithm.

In this study, absolute PCA loadings were used to identify the most important features. For each Principal Component, features with the highest absolute loadings were retained, as these loadings indicate the features that contribute most to the variance captured by each component. This approach preserves individual features with the strongest predictive power, allowing the model to remain interpretable—unlike using principal components directly, which could obscure relationships between original features and the outcome (20, 21). Selecting features based on PCA loadings is less common than using the components themselves but is especially beneficial when interpretation is a priority, as is often the case in healthcare applications (22, 23). By retaining features with the highest PCA loadings, the resulting model maintained the biological interpretability of individual biomarkers, which is crucial in clinical settings.

Subsequent to PCA, bivariate analysis was performed to explore the relationships between the outcome variable and each potential covariate and treatment variable. This was followed by stepwise regression and the application of the Random Forest algorithm, further refining the feature set to approximately 72 variables, thereby reducing noise and potential collinearity. The DML analysis was then performed on this refined dataset. This structured approach balanced model performance with interpretability, ensuring that the retained features provided diverse, biologically relevant information, minimized redundancy, and reduced collinearity—factors critical for enhancing model stability and performance.

2.3.4 *Confounders and covariates*

In terms of the DML framework, all features in this final list were used as both covariates and treatments. This approach is rooted in DML's robustness with high-dimensional data and its capacity to handle complex inter-variable relationships (15). Using all selected features as covariates and treatments ensured that confounders were accounted for, minimizing bias in the estimation of treatment effects. This aligns with a key premise of DML: its ability to manage high-dimensional datasets and correct for confounders to provide unbiased estimates. By including all relevant features in the model,

we reduced the risk of missing confounders that might otherwise distort the results.

2.4 Model development

The model development process in this study was guided by the need to accurately capture the relationships between treatment variables and ALS progression. This section outlines the steps taken to ensure robust model performance, including the application of Double Machine Learning (DML) techniques and the subsequent interpretation of results. By utilizing machine learning models and statistical frameworks, the aim was to not only predict outcomes but also identify and isolate true drivers of ALS progression. The development phase was divided into two main components: the application of DML for statistical causal inference and the interpretation of the results using a Mixed-Effects Linear Regression Framework. Each phase was carefully designed to minimize bias and ensure that the findings were generalizable.

2.4.1 *Double Machine Learning*

Once the dataset was prepared and key variables were defined, the model development phase began. The data was split into development (80%) and hold-out (20%) sets to ensure that the model could be validated on unseen data, thus minimizing the risk of overfitting. The Double Machine Learning (DML) algorithm was then applied, utilizing different machine learning models depending on the nature of the treatment and outcome variables. For implementing DML, the EconML package (24) from Microsoft Research was used in Python.

Double Machine Learning models are relatively new in literature, with the framework first formalized by Chernozhukov et al. (15). Their work introduced the orthogonalization technique, which built on prior methods in semiparametric statistics and econometrics but represented the first formal application of DML in high-dimensional machine learning contexts. This technique helps ensure the robustness of treatment effect estimates by effectively separating the influences of control variables from those of treatment variables. This approach is particularly well-suited for high-dimensional settings, where nuisance parameters—unrelated variables that could otherwise introduce bias—are accounted for to improve the accuracy of inference estimates.

The DML framework estimates two key components: the treatment models and the outcome models. The outcome variable for the study was defined as ALSFRS decline. For the outcome effect, a Random Forest Regressor model was utilized. The type of algorithm used for treatment models differed based on the label; for continuous treatments Random Forest Regressor algorithm, and for binary treatments XGBoost Classifiers were employed. These models were trained to estimate the nuisance parameters, which were then used to perform residual-based adjustments. This adjustment process de-biased the estimation of the treatment effect, allowing for robust results. The primary objective at this stage was to ensure that these nuisance components were accurately estimated so that the final model would be able to isolate the statistical causal relationships.

Hyperparameter optimization was performed using a grid search method, systematically evaluating different combinations of parameters to maximize model accuracy. This

approach aligned with the findings of Bach et al. (25), who emphasized the importance of hyperparameter tuning and variable selection in enhancing model performance in DML. Mean Square Error (MSE) was calculated for the treatment-to-outcome models to assess performance. For the confounder-to-treatment models, MSE was used when the treatment variable was continuous, and performance metrics such as Area Under the Curve (AUC), precision, and recall were employed when the treatment variable was binary. While performance metrics such as MSE and AUC were tracked to ensure model accuracy, the orthogonalization technique in DML ensured that the treatment effect estimates remained robust even when the covariate models were misspecified, as noted by Chernozhukov et al. (15). This served to mitigate the bias introduced by nuisance parameters.

Additionally, linear regression was applied to the residuals of both the treatment and outcome models to further refine the treatment effect estimates. These residual regressions helped eliminate any remaining bias in the treatment effect estimation. Finally, treatment effect estimation to possible external intervention was performed using these adjusted estimates, providing inferential results with associated confidence intervals and p-values.

2.4.2 *Interpretation framework*

The final phase of the methodology involved interpreting the results using a Mixed-Effects Linear Regression Framework. This framework was employed to analyze patient outcomes while extending the findings to broader factors, such as the impact on patient survival time. The mixed-effects model accounts for both fixed effects (the general progression of ALS across all patients) and random effects (variations between individual patients), allowing for a more detailed understanding of how treatments influence disease progression.

By integrating the results from the Double Machine Learning (DML) analysis with the mixed-effects model, the study translated treatment effects into practical, clinically relevant outcomes. Specifically, the coefficient of -0.022 from the mixed-effects model—representing the daily decline in ALSFRS scores—served as a multiplier to quantify how each unit change in the treatment effect slowed down ALS progression. This approach enabled the interpretation of the treatment's impact on ALSFRS scores in terms of days of slowed progression or extended survival time.

2.4.3 *In-Depth exploration of Double Machine Learning in causal inference*

The process of estimating treatment effects using Double Machine Learning (DML) is a critical component of this study, as it allows for robust inferences regarding the relationships between biomarkers and ALS progression. In this section, the Double Machine Learning framework is further elaborated.

The Double Machine Learning (DML) framework, as illustrated in Figure 3, operates in a two-step process designed to accurately estimate treatment effects in observational studies while accounting for high-dimensional confounders. This framework is particularly effective in estimating the treatment effects on ALS progression, controlling for various biomarkers and clinical factors that could introduce bias.

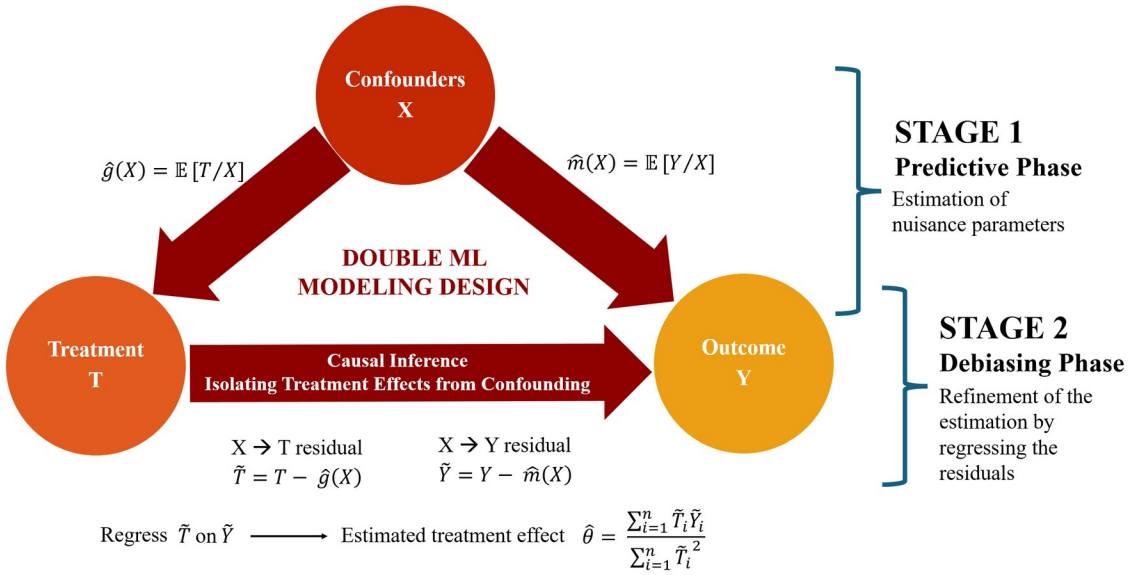

Figure 3. Visual representation of the Double Machine Learning (DML) framework: Stage 1 involves the estimation of nuisance parameters using machine learning models, which predict both the treatment (T) and outcome (Y) based on the covariates (X). The final treatment effect is derived from regressing the residuals, thereby eliminating the influence of confounders (X) on the treatment estimation in Stage 2.

In Stage 1, the nuisance parameters are estimated by regressing the confounders (X) on both the outcome (Y) and treatment (T). The goal of this step is to capture and model the relationships between the confounders, outcome, and treatment without introducing parametric assumptions.

The *outcome model* estimates the conditional expectation of the outcome given the covariates, represented as (1). This equation predicts the outcome Y based on the covariates X. $\hat{m}(X) = \mathbb{E}[Y/X]$ (eq1).

Similarly, the *treatment model* estimates the conditional expectation of the treatment given the covariates, represented by equation (2), which predicts the treatment T based on the covariates X. In the equations, $\hat{m}(X)$ and $\hat{g}(X)$ are essential components of the first stage. $\hat{g}(X) = \mathbb{E}[T/X]$ (eq2).

In Stage 2, the calculation from the first stage are used to debias the estimation. This begins with computing the residuals from the first-stage predictions, as shown in equations (3) and (4) as

$\hat{Y} = Y - \hat{m}(X)$ (eq3)

$\hat{T} = T - \hat{g}(X)$ (eq4)

These residuals represent the parts of the outcome and treatment that are not explained by the covariates X. The key idea is to estimate the treatment effect by regressing the residualized outcome on the residualized treatment using the formula given in (5)

$$\hat{\theta} = \left(\sum_{i=1} \tilde{T}_i \tilde{Y}_i\right) / \left(\sum_{i=1} \tilde{T}_i^2\right), \text{ where } i=1 \text{ to } n \quad \text{(eq5)}$$

where $\hat{\theta}$ is the estimated effect of T on Y after accounting for the confounders. Figure 3 illustrates this entire two-step regression process, showing how the Double ML framework effectively isolates the treatment effect from potential confounders.

This two-step process ensures that any confounding influence from X is removed, yielding an unbiased estimate of the effect of T on Y. To ensure robustness in estimating the treatment effect, the framework relies on the Neyman orthogonality property, which ensures that small errors from machine learning models in the first step do not bias the estimate. This principle ensures that the estimation is not heavily influenced by small changes in the estimation of the control variables, making the method robust to the choice of machine learning models and their associated biases in the first step.

To clarify the application of the Double ML framework in this research, it is essential to first define the key variables involved. As illustrated in Figure 4, in this study, the treatment variable (T) refers to the levels of biomarkers, such as calcium, potassium, and chloride, which are being studied to determine their impact on ALS progression. The outcome variable (Y) represents the progression of ALS, which is measured by changes in the ALSFRS score, reflecting the patient's functional decline over time. In this study, the confounders include patient demographics, clinical factors, and treatment history. Clinical factors encompass medical or health-related variables such as the type of ALS onset (bulbar *vs.* limb), and previous treatments the patient has undergone, and medical issues patient had previously. Other relevant confounders include vital signs and lab results that provide insights into the patient's health status.

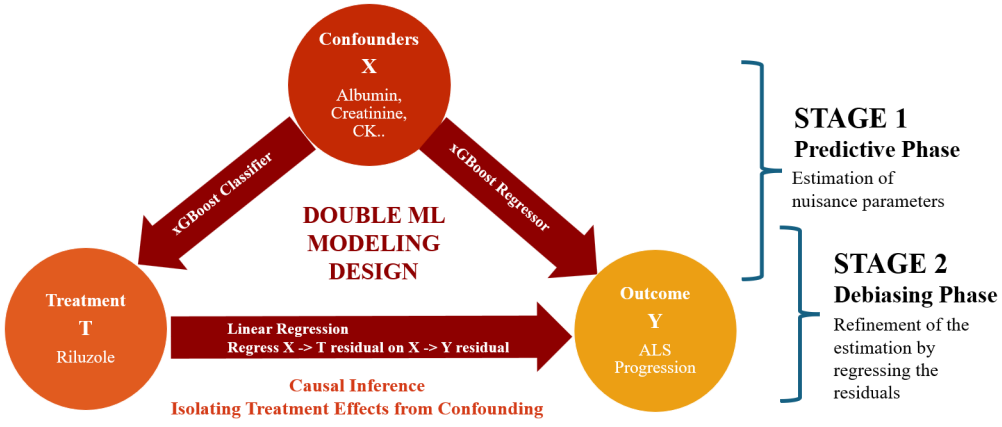

Figure 4. Double Machine Learning (DML) framework applied to ALS research with Riluzole usage. The figure illustrates the two-stage DML process to isolate the treatment effect on ALS progression. In Stage 1 (Predictive Phase), confounders such as Albumin, Creatinine, and Creatine Kinase (CK) are regressed on both the binary treatment variable (Riluzole usage) using an XGBoost Classifier and on the outcome (ALS progression) using an XGBoost Regressor. In Stage 2 (Debiasing Phase), the residuals from the predictive phase are then regressed to refine the estimation of the treatment effect of Riluzole usage on ALS progression, as measured by changes in ALSFRS scores. This approach helps to isolate the treatment effect from confounding influences, improving accuracy and reliability.

In Stage 1, an XGBoost Regressor is used to predict ALS progression, while an XGBoost Classifier is employed for Riluzole usage as it is a binary treatment. For continuous treatments, a Random Forest Regressor is applied. These models are crucial for capturing the relationships between confounders and both the treatment and the outcome. The chosen approach aligns with Chernozhukov et al.'s (15) recommendation to utilize flexible machine learning methods, such as Random Forests, Lasso, XGBoost, and others, capable of modeling complex relationships within the data. By adhering to this approach, the study ensures robust and accurate estimation of nuisance parameters in the Double Machine Learning (DML) framework, enhancing the validity of the causal inference insights.

In Stage 2, residuals from the first stage are computed to remove the effects of confounders using linear regression. This method, for example, allows the study to accurately estimate how Riluzole impacts ALS progression while controlling for the influence of covariates such as albumin, creatinine, and creatinine kinase.

Understanding complex diseases like ALS requires more than just recognizing patterns; it requires uncovering the underlying biomarkers that drive disease progression. Figure 5 presents a comparison between traditional Machine Learning (ML) and Double Machine Learning (DML) within the context of healthcare, specifically in understanding ALS progression. Machine Learning is depicted as a robust subset of AI, proficient in recognizing patterns and forecasting outcomes, with the goal of discerning correlations between biomarkers and ALS progression. It learns from complex patterns in data to enhance predictive modeling and flag potential markers.

On the other hand, Double Machine Learning is portrayed as a more advanced approach that integrates traditional econometric methods with machine learning techniques. This methodology focuses on identifying statistical effects rather than mere correlations, thereby providing insights that may help inform strategies for influencing the course of ALS progression. This shift from simple correlation to effect estimation enables more reliable inferences, which are crucial for designing interventions aimed at potentially altering ALS progression.

This study utilized all features that demonstrated predictive power in predicting ALS progression outcomes as both covariates and confounders within the analysis framework. This approach was grounded in the necessity to ensure robust and unbiased estimates of the treatment effects of interest. By treating these predictive features as both covariates and confounders, a wide range of potential sources of bias that could confound the relationship between treatment and outcome were controlled. This methodology aligns with the framework outlined by Chernozhukov et al. (15), which emphasizes the importance of incorporating predictive features in both treatment and outcome models to mitigate bias and ensure the robustness of statistical causal estimates through orthogonalization and cross-fitting techniques.

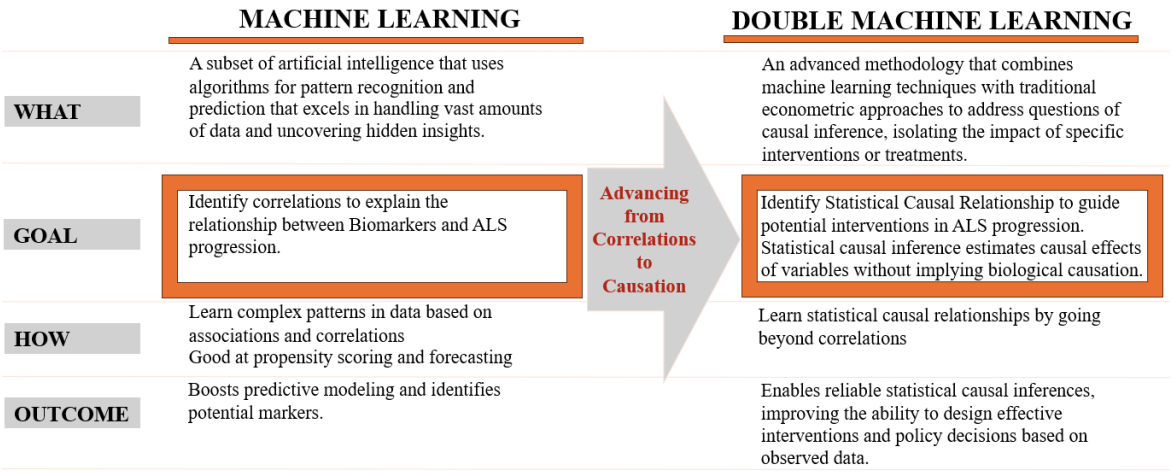

Figure 5. Comparison of Machine Learning (ML) and Double Machine Learning (DML) approaches in the context of ALS progression: While traditional ML focuses on identifying correlations to explain relationships between biomarkers and ALS progression, DML goes further by estimating treatment effects in a statistically rigorous way, providing insights that may inform targeted interventions.

3. Results

3.1 Overview of key findings

Table 1. Double Machine Learning Results on ALS Score Decline with Key Biomarker and Clinical Feature Insights: Summary of treatment effects for key features on ALSFRS score decline is reported. Results include estimated treatment effects, p-values, and 95% confidence intervals for each feature. Patients with gastrointestinal issues are identified based on hospital admissions or health records documenting a range of gastrointestinal conditions, as detailed in Note 1.

	Treatment Effect		Confidence Interval	
Feature	**Effect**	**p-value**	**Lower**	**Upper**
Bulbar onset	1.84	0.000	1.21	2.46
Gastrointestinal Issues[1]	1.19	0.000	0.69	1.70
Albumin	-0.18	0.001	-0.29	-0.08
Starting ALSFRS score	-0.37	0.000	-0.42	-0.32
Bicarbonate	0.60	0.000	0.42	0.77
Calcium	8.13	0.000	3.97	12.29
Chloride	-0.39	0.000	-0.51	-0.27
Creatine Kinase	-0.01	0.008	-0.02	0.00
Creatinine	-0.05	0.000	-0.06	-0.03
Creatinine % Change	-0.03	0.002	-0.04	-0.01
Glucose	0.27	0.016	0.05	0.49
Lymphocytes	-0.07	0.004	-0.12	-0.02
Phosphorus	6.06	0.000	3.35	8.78
Days since symptom onset	0.01	0.000	0.01	0.01
Symptom Onset in last 12 months	186.9	0.017	33.1	340.6

[1]Patients admitted to the hospital or with health records indicating any of the following conditions are marked as having gastrointestinal issues: abdominal hernias, anal and rectal conditions, benign and malignant neoplasms, dental and gingival conditions, diverticular disorders, exocrine pancreas conditions, gastrointestinal hemorrhages, inflammatory conditions, motility and defecation disorders, stenosis and obstruction, ulceration and perforation, vascular conditions, malabsorption disorders, oral soft tissue conditions, peritoneal and retroperitoneal issues, salivary gland conditions, and tongue conditions.

Based on the results from the Double Machine Learning (DML) model applied to ALS score decline, several biomarkers, including albumin, bicarbonate, calcium, chloride, creatine kinase, creatinine, creatinine change, glucose, lymphocytes, and phosphorus, demonstrated significant treatment effects on ALS progression. Additionally, clinical factors such as bulbar onset, gastrointestinal issues, starting ALSFRS score, and days since symptom onset were identified as other drivers of the ALSFRS score decline rate. These findings are summarized in Table 1, which consolidates key insights, while Figure 6 illustrates the confidence intervals of the treatment effects.

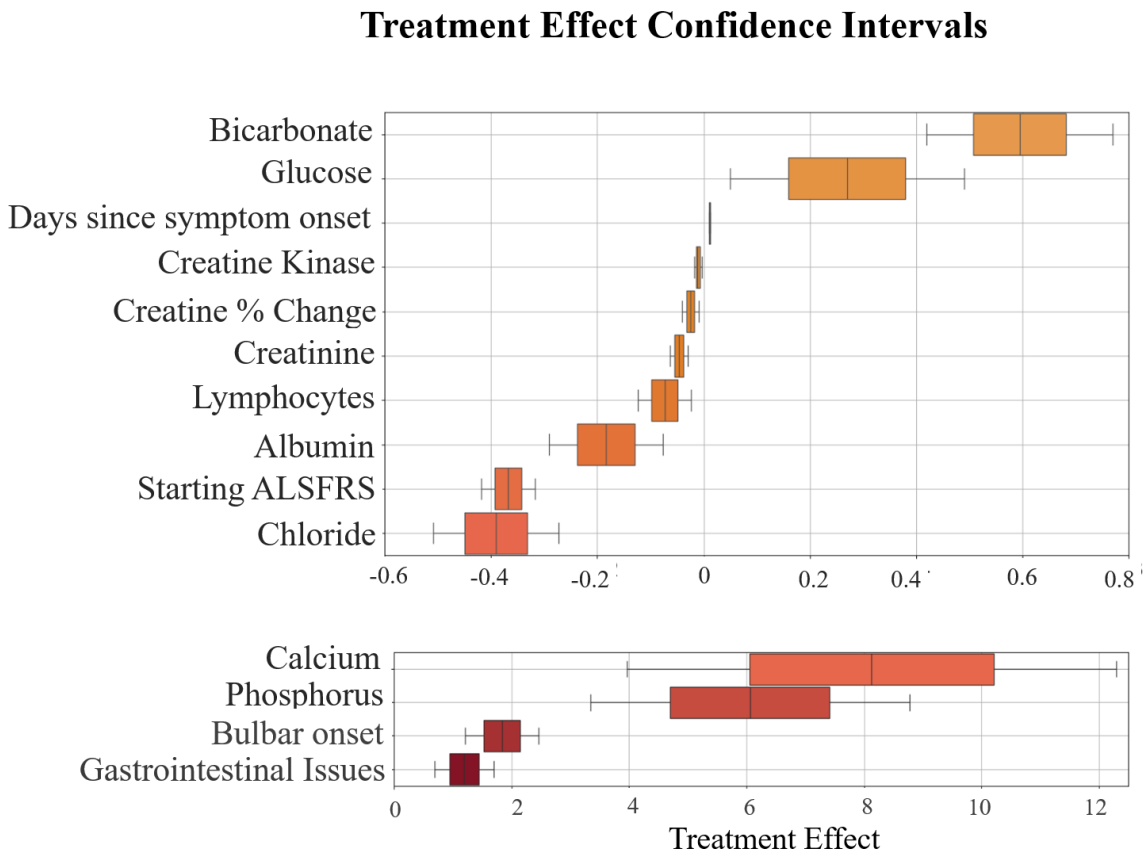

Figure 6. Confidence intervals for each treatment effect are displayed using the boxplot visuals. None of the features displayed has 0 in the confidence interval which indicates the effect estimations to be reliable.

The results reveal a clear distinction between the subset of correlated biomarkers and clinical factors that either accelerate or slow down ALS progression, with varying magnitudes of impact. Out of the many biomarkers tested, only a specific set, including albumin, bicarbonate, calcium, chloride, creatine kinase, creatinine, glucose, lymphocytes, and phosphorus, were identified as having significant effects on ALS progression.

Calcium showed the largest positive effect on ALS progression, with a treatment effect of 8.13, making it a significant contributor to faster disease progression. Similarly, symptom onset within the last 12 months and bicarbonate also exhibited strong positive effects, with treatment effects of 186.9 and 0.60, respectively, highlighting their association with rapid disease progression. Glucose and bulbar onset both showed notable positive impacts, with glucose having a treatment effect of 0.27 and bulbar onset at 1.84, suggesting their role in accelerating ALS progression, though at a smaller scale when compared with calcium.

In contrast, some factors exhibited a negative effect, suggesting a protective role in ALS progression. Chloride presented with one of the largest negative impacts, with a treatment effect of -0.39, followed by albumin, which also demonstrated a protective effect with a treatment effect of -0.18. Creatinine and lymphocytes also showed protective effects, with treatment effects of -0.05 and -0.07, respectively, though their influence was less pronounced. Creatinine percent change and creatine kinase showed minor protective effects, with treatment effects of -0.03 and -0.01, respectively, indicating their smaller-scale roles in slowing disease progression.

3.2 Biomarker analysis

One of the most notable findings was the relationship between serum albumin levels and ALS progression. The results showed a significant relationship between albumin levels and ALS progression. Higher serum albumin levels were associated with a slower decline in ALS scores, as reflected by the negative treatment effect of -0.18 (95% CI: -0.29 to -0.08, $p = 0.001$). Given that albumin is a marker of nutritional status, this finding suggests that better nutritional health may slow ALS progression. Albumin levels also inversely correlate with inflammatory cytokines; which are increased in ALS. Neuroinflammation has been shown to be associated with ALS pathogenesis.

The study also highlights the role of electrolyte imbalances in ALS. Elevated bicarbonate levels were associated with faster disease progression, with a treatment effect of 0.60 (95% CI: 0.42 to 0.77, $p < 0.001$). This association may indicate that respiratory dysfunction, which can increase bicarbonate levels, is also linked with faster progression. This association may reflect a complex, circular relationship: as ALS progresses, respiratory function may decline, leading to elevated bicarbonate levels, which in turn could accelerate disease progression. While the DML approach strengthens our ability to control for confounders, it cannot fully disentangle this potential feedback loop. Further investigation, including controlled studies, are needed to clarify the directional influence in this association.

Similarly, high calcium levels showed a strong association with rapid ALS progression (treatment effect = 8.13, 95% CI: 3.97 to 12.29, $p < 0.001$), suggesting a role for calcium dysregulation in ALS pathology. In contrast, higher chloride levels appeared to have a protective effect, with a treatment effect of -0.39 (95% CI: -0.51 to -0.27, $p < 0.001$), reinforcing the importance of monitoring and managing electrolytes. Ion channels, ion transporters and ionic homeostasis has been shown to influence the progression of ALS. Ca^{+2} is involved in synaptic transmission and

plasticity, as well as in the mitochondrial regulation of metabolic CNS pathways. Decreased Cerebro Spinal Fluid (CSF) Cl^{-1} levels are correlated with ALS progression. It may be speculated that a greater Cl^{-1} blood level is a compensatory mechanism to keep CSF Cl^{-1} levels elevated by increasing the concentration gradient from blood to CSF.

In addition, creatinine, a marker of muscle mass and kidney function, showed a protective effect, with higher levels linked to slower disease progression (treatment effect = -0.05, 95% CI: -0.06 to -0.03, $p < 0.001$). Elevated glucose levels, on the other hand, were associated with faster ALSFRS score decline (treatment effect = 0.27, 95% CI: 0.05 to 0.49, $p = 0.016$). Other significant biomarkers included creatine kinase and lymphocytes, both showing protective effects, while higher phosphorus levels were linked to faster disease progression.

3.3 Clinical factors influencing ALS progression

Among clinical factors, bulbar onset was found to significantly accelerate ALS progression. Patients with bulbar-related symptoms experienced faster disease progression, reflected by a treatment effect of 1.84 (95% CI: 1.21 to 2.46, $p < 0.001$). This confirms the severity of bulbar onset ALS, which typically results in difficulties with speech and swallowing, leading to a more aggressive disease course. Gastrointestinal issues were also identified as contributing to faster disease progression, with a treatment effect of 1.19 (95% CI: 0.69 to 1.70, $p < 0.001$). This finding suggests that monitoring gastrointestinal health is a relevant component of ALS management and could influence treatment strategies. Although bulbar-related symptoms and gastrointestinal issues are typically covariates, they were analyzed as treatment variables in this study to explore their direct impact on ALS progression. This approach allowed for assessing of how bulbar symptoms influence disease severity, providing important insights for ALS management.

The initial ALSFRS score emerged as another crucial determinant of disease progression. A higher starting ALSFRS score, which indicates better initial motor function, was linked to a slower progression of the disease (treatment effect = -0.37, 95% CI: -0.42 to -0.32, $p < 0.001$). This finding reinforced the value of early and accurate functional assessments in predicting the disease trajectory. It highlights the importance of starting interventions early when a patient's functional abilities are still relatively preserved, allowing for more effective patient care strategies. Another possible explanation is that as patients' ALSFRS scores decrease, reflecting more advanced stages of the disease, their functional decline tends to accelerate (26, 27). Lower ALSFRS scores are associated with a faster rate of decline, possibly due to increased complications such as respiratory failure or bulbar dysfunction.

Patients who experienced their first ALS symptoms within the last 12 months exhibited a large treatment effect of 186.9 (95% CI: 33.1 to 340.6, $p = 0.017$), indicating a rapid disease progression. The wide confidence interval suggests that the actual effect could vary. However, this finding implies that recent symptom onset is a critical predictor of disease progression, potentially signaling a more aggressive form of ALS. A plausible

explanation for this rapid decline is that patients whose symptoms began within the last 12 months are likely experiencing a more aggressive form of ALS. ALS is very difficult to diagnose, with many patients going undiagnosed for years despite showing early symptoms. The fact that these patients were diagnosed within a short time frame—within 12 months of symptom onset—may indicate a more severe and fast-progressing form of the disease, where symptoms become apparent and debilitating more quickly. As a result, these patients experience a steeper decline in their ALSFRS scores. This insight emphasizes the need for heightened monitoring and possibly more intensive interventions for patients who have experienced recent symptom onset. By treating the timing of symptom onset as a treatment variable in this analysis, the study demonstrated its significant impact on ALS progression. Understanding this relationship could guide clinicians to tailor intervention strategies for patients in the early stages of ALS, ensuring that care is administered proactively during these crucial months.

3.4 Non-causal variables

Although some biomarkers and clinical factors were initially correlated with ALS progression in preliminary machine learning models, further analysis using DML did not establish them as statistically significant predictors with a direct effect on disease progression. Factors such as creatine kinase percent change, white blood cell change, age, hematocrit, platelets, red blood cells, phosphorus percent change, absolute eosinophil count, alkaline phosphatase, and Riluzole usage did not show a measurable impact on ALS progression in this context. Notably, Riluzole, the most prescribed medication for ALS, was found to have no statistically significant effect on disease progression, consistent with prior studies (28).

3.5 Insights from mixed-effects modeling

To further enhance the interpretability of the treatment effect, a mixed-effects model was used to regress ALSFRS scores against the number of days since the scores were recorded. A mixed-effects model, also known as a hierarchical or multilevel model, is particularly useful in situations where data is structured in a way that there are both fixed effects (which apply to all individuals or units in the study) and random effects (which account for individual variations or repeated measures within the same individual). In this study, the mixed-effects model was appropriate because ALSFRS scores were measured repeatedly over time for the same patients, leading to potential correlations within each patient's data. The model allowed us to account for these within-subject correlations while also estimating the overall treatment effects.

This regression analysis provided quantitative estimates that offered critical insights into potential life expectancy extensions for patients receiving treatment. The mixed-effects regression yielded a coefficient of -0.022 for the variable “days” (slope of ALSFRS decline over time), representing the rate of decline in ALSFRS score per day. This unitless coefficient indicated an average daily decrease of 0.022 points in ALSFRS, reflecting a gradual decline in function as ALS progressed.

Upon applying the mixed-effects model coefficient to the treatment effects of blood markers, several insights emerged. For instance, increasing Albumin levels by 0.25 mmol/L (normal range: 0.5-0.75 mMol/L) led

to a 0.046-point deceleration in FRS score progression, correlating with an 2.1 day increase in life expectancy. Conversely, decreasing Bicarbonate levels by 10 mmol/L (normal range: 22-32 mMol/L) result in a 5.95-point acceleration in FRS score progression, calculationg to a 270 day increase in life expectancy. Similarly, decreasing Calcium levels by 0.5 mmol/L (normal range: 2.2-2.7 mMol/L) calculated with an 4.1-point decrease in FRS score progression, translating to an increase of 184 days in life expectancy. Decreasing Phosphorus levels by 0.25 mmol/L (normal range: 1-1.5 mMol/L), calculated to an increase of 69 days in life expectancy on average. In contrast, increasing Chloride levels by 5 mmol/L (normal range: 96-106 mMol/L) calculated to a 1.95-point slowdown in FRS score progression, leading to an 88.6 day increase in life expectancy. Furthermore, elevations in Creatine Kinase by 1 U/L unit and Creatinine by 1 mmol/L unit result in respective reductions in FRS score progression, leading to increases in life expectancy of 0.5 days and 2 days, respectively. Lastly, decreasing Glucose levels by 0.5 mmol/L were calculated with a 0.135-point deceleration in FRS score progression, resulting in an increase of 6 days in life expectancy. While such calculations obviously are not algorithmically compressible, it is possible that rigorous control of common blood markers such as glucose, ions and electrolytes can significantly (of the order of months-to-years) increase the lifespan of ALS patients.

In summary, calcium and glucose have the largest positive effects, driving faster ALS progression, while creatinine and lymphocytes offer protective effects but on a smaller scale. However, creatine kinase's effect, although negative, is minimal compared to other factors such as chloride and albumin, which have a stronger protective impact on slowing ALS progression.

It is important to note that these effects should be interpreted within the context of typical biomarker ranges; for example, even a 1 mmol/L change is substantial when compared to standard variability; which is why levels were adjusted based on their normal ranges in the blood. Normalizing these effects by common ranges can provide additional insight into the relative impact of each biomarker.

4. Discussion

This study applies advanced Double Machine Learning (DML) techniques to identify key biomarkers that influence ALS progression, providing a novel approach to understanding statistical causal relationships underlying the disease. Leveraging the DML framework, this research offers critical new insights into ALS progression by estimating the treatment effects of various biomarkers. Significant biomarkers identified in this study include bicarbonate, glucose, creatinine, creatine kinase, lymphocytes, albumin, phosphorus, chloride, and calcium, each potentially playing a role in disease progression.

The findings of this study build on prior research, providing a more nuanced understanding of how specific biomarkers may statistically impact ALS progression. For instance, while previous studies identified correlations between white blood cell counts and ALS progression, applying DML in this analysis offers a more rigorous statistical assessment, allowing for insights that suggest potential causal associations. In alignment with

Murdock et al. (29), who identified correlations between total white blood cell counts in ALS patients, this study clarifies the specific role of lymphocytes, showing that higher lymphocyte levels are associated with slower ALS progression. This finding contrasts with Yang et al. (30), who reported no correlation between lymphocyte levels and ALS progression, despite observing statistically significant differences in lymphocyte levels in ALS patients.

Electrolytes, particularly chloride, bicarbonate, and calcium, also play crucial roles in ALS progression. While previous studies (31-34) identified correlations between these electrolytes and ALS outcomes, this study leveraged DML to show that elevated bicarbonate and calcium levels accelerated ALS progression, while higher chloride levels were protective. Mineral balance, particularly calcium and phosphorus, further emerged as a critical factor in ALS progression. Previous research (35, 36) explored their roles, but the DML framework applied here provided stronger evidence that higher levels of these minerals drove faster disease progression.

These findings underscore the importance of monitoring blood electrolytes and mineral levels in ALS patients, as they may serve as valuable therapeutic targets. This perspective was supported by Sirabella et al. (37), who emphasized that disruptions in ionic homeostasis—including critical ions such as calcium, sodium, and iron—played a significant role in ALS pathophysiology. Their work highlighted the potential of targeting ionic channels and transporters to restore balance as a therapeutic approach. Taken together, these insights suggest that a strategy of carefully regulating electrolyte levels, particularly bicarbonate, calcium, Phosphorous and chloride, could offer promising avenues for managing ALS progression.

The protective roles of creatinine and creatine kinase (CK) were also reinforced in this study. Previous studies (38, 39) identified these biomarkers as significant, and this study confirmed their impact, demonstrating that both creatinine and CK slowed ALS progression. Similarly, Ahangaran et al. (40), using a probabilistic causal discovery model, found a significant effect of CK on ALS progression.

Higher blood glucose levels, which Sun et al. (38) associated with increased mortality, and elevated cerebrospinal fluid glucose levels, correlated with faster ALS progression (41, 42). This study confirmed the significant role of glucose in ALS, specifically demonstrating that higher blood glucose levels contributed to faster ALS progression. These findings are consistent with those of Ahangaran et al. (40), who also identified a causal association between glucose levels and ALS progression. Additionally, Lerskiatiphanich et al. (43) highlighted systemic disruptions in glucose metabolism in ALS, noting increased glucose uptake and glycogen accumulation, alongside glucagon resistance, in the presence of ALS, further supporting the link between impaired glucose homeostasis and ALS progression.

The role of albumin in ALS has been explored in several studies (38, 39, 44, 45) with varying conclusions. Hertel et al. (45) found a correlation between albumin levels and ALSFRS scores but did not establish a significant link to disease progression. In

contrast, Sun et al. (38) and Gentile et al. (39) reported that lower albumin levels were associated with reduced survival in ALS patients. Additionally, Chiò et al. (44) showed higher albumin was linked to better survival outcome. This study provided further evidence that higher albumin levels are associated with slower ALS progression, offering a more comprehensive understanding of albumin's protective role in the disease.

Non-blood markers, such as bulbar onset, were also reaffirmed as significant. Consistent with previous research (46, 47), this study confirmed that bulbar onset was linked to worse outcomes. Additionally, the timing of disease onset pre-diagnosis emerged as a significant contributor to disease progression. Despite the inclusion of variables such as ethnicity, age, and gender in the analysis, they did not emerge as significant factors influencing ALS progression in this study. Given the potential for these demographic factors to impact disease progression and treatment response in broader populations, future research should explore these variables in more detail. Larger datasets or more focused studies may reveal their role in modifying the effects of biomarkers or other clinical factors.

In the literature, several biomarkers have been associated with ALS progression. For example, studies by Murdock et al. (29) and Cui et al. (48) reported associations between changes in white blood cell count and ALS progression, while Mandrioli et al. (28) identified higher hematocrit levels as being linked to poorer outcomes. Additionally, Dupuis et al. (49) found platelets to be significant for ALS progression, and Gentile et al. (39) observed a negative correlation between potassium levels and ALSFRS-R scores. However, this study's DML analysis suggested that these variables (except for potassium) did not have a measurable impact on ALS progression when considering a statistical DML framework. Similarly, while Yang et al. (30) noted correlations with absolute eosinophil count, and Mandrioli et al. (28) suggested an effect of Riluzole in slowing disease progression, none of these factors were found to significantly influence ALS progression in our analysis.

These findings underscore the importance of distinguishing between correlation and statistical causal inference, as only statistical causal drivers can provide actionable insights for developing effective ALS interventions. The study by Wang (50) is the only one in the literature applying Double Machine Learning (DML) in the context of ALS research, specifically focusing on genomics data. While her work centers on identifying relationships among genomic features, our study diverges by focusing on blood biomarkers to identify drivers of ALS progression.

A key strength of this study was its comprehensive analysis, encompassing over 400 features and data from approximately 6,000 ALS patients—one of the largest datasets used in the field. The extensive feature set, including newly created longitudinal variables, captured the dynamic nature of ALS, offering deeper insights into the disease's complexity. This large cohort not only enhanced generalizability but also allowed for a nuanced understanding of how multiple biomarkers interact to drive disease progression, potentially uncovering significant biomarkers that may have been overlooked in smaller studies.

However, several limitations must be considered. DML can be affected by collinearity and interactions between features, and it can be sensitive to overfitting and model specification. While DML is robust in handling confounding in high-dimensional datasets, it cannot fully account for residual confounding, particularly when certain variables are unmeasured or missing. Examples include unmeasured blood biomarkers, CSF markers, genetic factors, and environmental influences. Additionally, imperfect measurement of blood biomarkers may introduce biases that cannot be entirely avoided. The reliance on the PRO-ACT dataset also introduces potential biases. The dataset, derived from clinical trials, may not fully represent the broader ALS population, leading to selection bias. Furthermore, the high percentage of missing data limits the generalizability of these findings. Anonymization constraints further restrict the ability to integrate important demographic or genetic factors into the analysis. Moreover, the intermittent collection of data points in the PRO-ACT dataset limited the recency of the analysis, as the dataset extended only until 2022, potentially missing more recent developments in ALS or treatment responses. Hence, the study is planned to be reiterated when new data becomes available, allowing for the inclusion of more recent findings and a more comprehensive analysis.

Despite these limitations, several steps were taken to mitigate these challenges. To address collinearity and interactions between features, tree-based models were employed in the predictive phase (stage 1). Tree-based models are well-suited to handle correlated features and complex interactions commonly found in high-dimensional biological systems like ALS. Additionally, as Fuhr et al. (17) demonstrated, DML's capacity to adjust for various confounding influences is enhanced when flexible machine learning models, such as boosted trees, neural networks, or random forests, are used and when the sample size is reasonably large. This study's use of tree-based models and a large cohort aligns with these findings, aiming to support the robustness of treatment effect estimates in the presence of confounding factors. To manage the high percentage of missing data in the PRO-ACT dataset, imputation techniques were applied to reduce the potential biases introduced by incomplete data.

However, despite these efforts, unobserved confounders may still affect the results, potentially biasing causal estimates and limiting the study's conclusions. For future research, expanding the analysis to include more comprehensive data—such as additional biomarkers, genetic information, and environmental factors—could help identify more confounders and further reduce bias. Additionally, conducting prospective clinical trials in more diverse ALS populations will be crucial for validating these findings and ensuring that the results are generalizable.

5. Conclusion

This research sought to advance the understanding of ALS progression through the application of advanced analytical models, specifically Double Machine Learning (DML), a novel approach in ALS biomarker studies. To our knowledge, DML has not been previously applied to ALS blood biomarkers, making this study a significant contribution to the field by providing new insights into the statistical causal relationships between biomarkers and

ALS progression. The analysis identified several significant biomarkers, including bicarbonate, glucose, creatinine, creatine kinase, lymphocytes, albumin, phosphorus, chloride, and calcium, all of which were found to play critical roles in determining ALS progression.

While clinical trials remain the gold standard for establishing causality, they often face challenges in fully accounting for patient differences and the complexities of real-world scenarios. Given the low prevalence of ALS, conducting large-scale experimental studies is inherently challenging. In this context, the insights gained from DML can help clinical trials to narrow down the most relevant biomarkers to focus on, thereby enhancing study efficiency and potentially accelerating the identification of effective therapeutic targets. By identifying features with significant treatment effects on ALS progression, DML provides a strategic roadmap for more targeted and efficient therapeutic strategies.

Additionally, the integration of mixed-effects models enhances the temporal analysis of ALS progression. The combination of DML with mixed-effects modeling represents a novel approach that can serve as a template for future studies aiming to explore complex statistical relationships in high-dimensional datasets in neurodegenerative research.

While the DML approach provided promising insights into ALS progression, validation through prospective clinical trials is essential to confirm these findings, as the relationships identified here were based on retrospective, observational data. In addition, prospective clinical trials in diverse ALS populations will be critical to validate these findings and ensure their generalizability.

Future research should aim to include more comprehensive data, such as additional biomarkers, genetic information, and environmental factors. The use of other advanced statistical causal inference methods, such as the Deep End-to-End Causal Inference (DECI) model (51), can further explore the dynamic interactions between biomarkers and the sequencing of events among biomarkers as they change over time, potentially leading to improved therapeutic interventions for ALS patients.

In conclusion, this study demonstrated the power of DML to uncover statistical causal relationships between biomarkers and ALS progression, utilizing an extensive number of features and a large patient sample. This approach provided robust and precise insights into factors that may influence ALS progression, offering a strong foundation for informing future intervention strategies.

Acknowledgements

I would like to extend my sincere gratitude to Ms. Kate Allender from Tesla STEM High School for her unwavering support and encouragement. Her guidance was instrumental in shaping my academic and personal growth. I am deeply thankful to my mentors, Professor Dr. Stephen Tapscott at University of Washington Neurology; Associate Professor Dr. Berkin Bilgic at Harvard Medical School Radiology; Dr. Emre Kiciman, Senior Principal Researcher at Microsoft; and Dr. Aurora Sáez from Córdoba University of Spain for their invaluable guidance and advice. Their expertise

and mentorship greatly contributed to the success of this project.

Data used in the preparation of this article was obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. 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: ALS Therapy Alliance, Cytokinetics, Inc., Amylyx Pharmaceuticals, Inc., Knopp Biosciences, Neuraltus Pharmaceuticals, Inc., Neurological Clinical Research Institute, MGH, Northeast ALS Consortium, Novartis, Prize4Life Israel, Regeneron Pharmaceuticals, Inc., Sanofi, Teva Pharmaceutical Industries, Ltd., The ALS Association.

References

1. Costa, J., De Carvalho, M. (2016). Emerging molecular biomarker targets for amyotrophic lateral sclerosis. *Clinica Chimica Acta*, *455*, 7–14. https://doi.org/10.1016/j.cca.2016.01.011

2. Thakore, N. J., Pioro, E. P., Udeh, B. L., Lapin, B. R., Katzan, I. L. (2020). A Cost-Effectiveness Framework for Amyotrophic Lateral Sclerosis, Applied to Riluzole. Value in Health, 23(12), 1543–1551. https://doi.org/10.1016/j.jval.2020.06.012

3. Mehta, P., Raymond, J., Zhang, Y., Punjani, R., Han, M., Larson, T., et al. (2023). Prevalence of amyotrophic lateral sclerosis in the United States, 2018. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, *24*(7–8), 702–708. https://doi.org/10.1080/21678421.2023.2245858

4. Arthur, K. C., Calvo, A., Price, T. R., Geiger, J. T., Chiò, A., Traynor, B. J. (2016). Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. *Nature Communications*, *7*(1). https://doi.org/10.1038/ncomms12408

5. Achtert, K., Kerkemeyer, L. (2021). The economic burden of amyotrophic lateral sclerosis: a systematic review. *The European Journal of Health Economics*, *22*(8), 1151–1166. https://doi.org/10.1007/s10198-021-01328-7

6. Tzeplaeff, L., Wilfling, S., Requardt, M. V., Herdick, M. (2023). Current state and future directions in the therapy of ALS. *Cells*, *12*(11), 1523. https://doi.org/10.3390/cells12111523

7. Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, *9*(8). https://doi.org/10.1098/rsos.220638

8. Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., et al. (2024b). Causal machine learning for predicting treatment outcomes. *Nature Medicine*, *30*(4), 958–968. https://doi.org/10.1038/s41591-024-02902-1

9. Peet, E. D., Schultz, D., Lovejoy, S., Tsui, F. (2022). Variation in the infant health effects of the women, infants, and children program by predicted risk using novel machine learning methods. *Health Economics*, *32*(1), 194–217. https://doi.org/10.1002/hec.4617

10. Ohnmacht, A. J., Avar, G., Schübel, M. K., O'Neill, T. J., Krappmann, D., Menden, M. P. (2024). The pharmacogenomic assessment of molecular epithelial-mesenchymal transition signatures reveals drug susceptibilities in cancer cell lines. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2024.01.16.575190

11. Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*(none). https://doi.org/10.1214/09-ss057

12. Russo, F., Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, *21*(2), 157–170. https://doi.org/10.1080/02698590701498084

13. Williamson, J. (2019). Establishing causal claims in medicine. *International Studies in the Philosophy of Science*, *32*(1), 33–61. https://doi.org/10.1080/02698595.2019.1630927

14. Khan, K. S., Ball, E., Fox, C. E., Meads, C. (2012). Systematic reviews to evaluate causation: an overview of methods and application. *Evidence-Based Medicine*, *17*(5), 137–141. https://doi.org/10.1136/ebmed-2011-100287

15. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., et al. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

16. PRO-ACT (Project for ALS Clinical Trials). *PRO-ACT - HOME*. (n.d.). Pooled Resource Open-Access ALS Clinical Trials Database. https://ncri1.partners.org/ProACT/Document/DisplayLatest/9

17. Fuhr, J., Berens, P., Papies, D. (2024, March 21). *Estimating Causal Effects with Double Machine Learning -- A Method Evaluation*. arXiv.org. https://arxiv.org/abs/2403.14385v1

18. Angrist, J. D., Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. https://doi.org/10.2307/j.ctvcm4j72

19. Cinelli, C., Forney, A., Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*, 004912412210995. https://doi.org/10.1177/00491241221099552

20. Chahboun, S., Maaroufi, M. (2021). Principal component Analysis and Machine learning approaches for photovoltaic power prediction: A Comparative study. *Applied Sciences*, *11*(17), 7943. https://doi.org/10.3390/app11177943

21. Wuttichaikitcharoen, P., Babel, M. (2014). Principal component and multiple regression analyses for the estimation of suspended sediment yield in ungauged basins of northern Thailand. *Water*, *6*(8), 2412–2435. https://doi.org/10.3390/w6082412

22. Habehh, H., Gohel, S. (2021). Machine learning in healthcare. *Current Genomics*, *22*(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

23. Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31 57. https://doi.org/10.1145/3236386.3241340

24. PyWhy. (2023). *EconML:* A Python package for econometrics and machine learning. EconML implements orthogonal machine learning algorithms such as the double machine learning work of Chernozhukov et al. This toolkit is designed to measure the causal effect of some treatment variable(s) t on an outcome variable y, controlling for a set of features x. GitHub repository. Retrieved from https://github.com/py-why/EconML

25. Bach, P., Schacht, O., Chernozhukov, V., Klaassen, S., Spindler, M. (2024). Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study. *PMLR*, 1065–1117. https://proceedings.mlr.press/v236/bach24a.html

26. Devos, D., Moreau, C., Kyheng, M., Garçon, G., Rolland, A. S., Blasco, H., et al. (2019). A ferroptosis–based panel of prognostic biomarkers for Amyotrophic Lateral Sclerosis. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-39739-5

27. Prell, T., Witte, O. W., Gunkel, A., Grosskreutz, J. (2020). Cognitive deficits have only limited influence on health-related quality of life in amyotrophic lateral sclerosis. *Aging & Mental Health, 24*(12), 1963–1967. https://doi.org/10.1080/13607863.2019.1642296

28. Mandrioli, J., Rosi, E., Fini, N., Fasano, A., Raggi, S., Fantuzzi, A. L., et al. (2017). Changes in routine laboratory tests and survival in amyotrophic lateral sclerosis. *Neurological Sciences*, *38*(12), 2177–2182. https://doi.org/10.1007/s10072-017-3138-8

29. Murdock, B. J., Zhou, T., Kashlan, S. R., Little, R. J., Goutman, S. A., Feldman, E. L. (2017). Correlation of peripheral immunity with rapid amyotrophic lateral sclerosis progression. *JAMA Neurology*, *74*(12), 1446. https://doi.org/10.1001/jamaneurol.2017.2255

30. Yang, J., Liu, T., Zhang, L., Li, X., Du, F. P., Liu, Q., et al. (2023). Eosinophils at diagnosis are elevated in amyotrophic lateral sclerosis. *Frontiers in Neurology*, *14*. https://doi.org/10.3389/fneur.2023.1289467

31. Stambler, N., Charatan, M., Cedarbaum, J. M. (1998). Prognostic indicators of survival in ALS. *Neurology*, *50*(1), 66–72. https://doi.org/10.1212/wnl.50.1.66

32. Qureshi, M., Shui, A., Dibernardo, A. B., Brown, R. H., Schoenfeld, D. A., Cudkowicz, M. E. (2008). Medications and laboratory parameters as prognostic factors in amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis*, *9*(6), 369–374. https://doi.org/10.1080/17482960802163614

33. Ong, M., Tan, P. F., Holbrook, J. D. (2017). Predicting functional decline and survival in amyotrophic lateral sclerosis. *PLoS ONE*, *12*(4), e0174925. https://doi.org/10.1371/journal.pone.0174925

34. Manera, U., Grassano, M., Matteoni, E., Bombaci, A., Vasta, R., Palumbo, F., et al. (2023). Serum chloride as a respiratory failure marker in amyotrophic lateral sclerosis. *Frontiers in Aging Neuroscience*, *15*. https://doi.org/10.3389/fnagi.2023.1188827

35. Katzeff, J. S., Bright, F., Lo, K., Kril, J. J., Connolly, A., Crossett, B., et al. (2020). Altered serum protein levels in frontotemporal dementia and amyotrophic lateral sclerosis indicate calcium and immunity dysregulation. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-70687-7

36. Gordon, J., Lerner, B. (2019). Insights into Amyotrophic Lateral Sclerosis from a Machine Learning Perspective. *Journal of Clinical Medicine*, *8*(10), 1578. https://doi.org/10.3390/jcm8101578

37. Sirabella, R., Valsecchi, V., Anzilotti, S., Cuomo, O., Vinciguerra, A., Cepparulo, P., et al. (2018). Ionic homeostasis maintenance in ALS: Focus on new therapeutic targets. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00510

38. Sun, J., Carrero, J. J., Zagai, U., Evans, M., Ingre, C., Pawitan, Y., et al. (2020). Blood biomarkers and prognosis of amyotrophic lateral sclerosis. *European Journal of Neurology*, *27*(11), 2125–2133. https://doi.org/10.1111/ene.14409

39. Gentile, F., Maranzano, A., Verde, F., Bettoni, V., Colombo, E., Doretti, A., et al. (2023). The value of routine blood work-up in clinical stratification and prognosis of patients with amyotrophic lateral sclerosis. *Journal of Neurology*, *271*(2), 794–803. https://doi.org/10.1007/s00415-023-12015-3

40. Ahangaran, M., Peimankar, A., Wiil, U. K., Chio, A. (2022). Impact of laboratory biomarkers on ALS disease progression: a probabilistic causation approach. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *11*(1). https://doi.org/10.1007/s13721-022-00386-y

41. Gray E, Larkin JR, Claridge TD, Talbot K, Sibson NR, Turner MR. (2015).The longitudinal cerebrospinal fluid metabolomic profile of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener*. 16(7-8), 456-63. https://doi.org/10.3109/21678421.2015.1053490

42. Wuolikainen A, Jonsson P, Ahnlund M, Antti H, Marklund SL, Moritz T, et al. (2016). Multi-platform mass spectrometry analysis of the CSF and plasma metabolomes of rigorously matched amyotrophic lateral sclerosis, Parkinson's disease and control subjects. *Mol Biosyst.* 12(4), 1287-98. https://doi.org/10.1039/c5mb00711a

43. Lerskiatiphanich, T., Lee, J., Marallag, J. (2022). Glucose metabolism in amyotrophic lateral sclerosis: it is bitter-sweet. *Neural Regeneration Research*, *17*(9), 1975. https://doi.org/10.4103/1673-5374.335154

44. Chiò, A., Calvo, A., Bovio, G., Canosa, A., Bertuzzo, D., Galmozzi, F., et al. (2014). Amyotrophic lateral sclerosis outcome measures and the role of albumin and creatinine. *JAMA Neurology*, *71*(9), 1134. https://doi.org/10.1001/jamaneurol.2014.1129

45. Hertel, N., Kuzma-Kozakiewicz, M., Gromicho, M., Grosskreutz, J., De Carvalho, M., Uysal, H., et al. (2022). Analysis of routine blood parameters in patients with amyotrophic lateral sclerosis and evaluation of a possible correlation with disease progression—a multicenter study. *Frontiers in Neurology*, *13*. https://doi.org/10.3389/fneur.2022.940375

46. Christensen, P. B., Højer-Pedersen, E., Jensen, N. B., MD. (1990). Survival of patients with amyotrophic lateral sclerosis in 2 Danish counties. *Neurology*, *40*(4), 600. https://doi.org/10.1212/wnl.40.4.600

47. Testa D, Lovati R, Ferrarini M, Salmoiraghi F, Filippini G. Survival of 793 patients with amyotrophic lateral sclerosis diagnosed over a 28-year period. Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders: Official Publication of the World Federation of Neurology,

The Amyotrophic Lateral Sclerosis Functional Rating Scale. (1996). *Archives of Neurology*, *53*(2), 141. https://doi.org/10.1001/archneur.1996.00550020045014

48. Cui, C., Ingre, C., Yin, L., Li, X., Andersson, J., Seitz, C., et al. (2022). Correlation between leukocyte phenotypes and prognosis of amyotrophic lateral sclerosis. *eLife*, *11*. https://doi.org/10.7554/elife.74065

49. Dupuis, L., Spreux-Varoquaux, O., Bensimon, G., Jullien, P., Lacomblez, L., Salachas, F., et al. (2010). Platelet serotonin level predicts survival in amyotrophic lateral sclerosis. *PLoS ONE*, *5*(10), e13346. https://doi.org/10.1371/journal.pone.0013346

50. Wang, C. (2021). The Application of Double Machine Learning Onto Genomics Data Associated with Amyotrophic Lateral Sclerosis. In Massachusetts Institute of Technology, *Massachusetts Institute of Technology*. https://dspace.mit.edu/bitstream/handle/1721.1/139444/Wang-cwang506-meng-eecs-2021-thesis.pdf?sequence=1&isAllowed=y

51. Geffner, T., Antoran, J., Foster, A., Gong, W., Ma, C., Kiciman, E., et al. (2022b). *Deep end-to-end causal inference*. arXiv.org. https://arxiv.org/abs/2202.02195