Individualized predictions for clinical milestone in amyotrophic lateral sclerosis: A multialgorithmic approach

Hyeon-Ji Oh[1], Won-Joon Lee[1], Jung-Joon Sung[2], Yoon-Ho Hong[3] (iD)
and the Pooled Resource Open-Access ALS Clinical Trials Consortium

Abstract

Objective: The phenotypic heterogeneity and complex disease trajectory complicate the ability to predict specific clinical milestone for individual patients with amyotrophic lateral sclerosis (ALS). Here we developed individualized prediction models to estimate the time to the loss of autonomy in swallowing function.

Methods: Utilizing the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database, we built three models of distinct time-to-event prediction algorithms: accelerated failure time (AFT), cox proportional hazard (COX) and random survival forest (RSF) for an individualized risk assessment of the swallowing milestone. The target variable was defined as the time to a decline in the ALSFRS-R swallowing item score to 1 or below, indicating a need for supplementary tube feeding.

Results: Internal cross-validation revealed the median concordance index (C-index) of 0.851 (IQR, 0.842–0.859) for AFT, 0.850 (0.841–0.859) for COX and 0.846 (0.839–0.854) for RSF, and all models demonstrated good distributional calibration with predicted and observed event probabilities closely matched across different time intervals. For external validation with a registry dataset with characteristics different from PRO-ACT, the discriminative power was replicated with comparable C-indices for all models, whereas the calibration revealed a left-skewed distribution suggesting a bias towards overestimation of event probabilities in real-world data. While all models were effective at stratifying patients, the results of RSF model, unlike AFT and COX, did not match well with the KM curves of the corresponding risk groups, supporting the importance of nuanced understanding of data structure and algorithmic properties.

Conclusion: Our models are implemented into a web application which could be applied to individualized counselling, management and clinical trial design for gastrostomy intervention. Further studies for model optimization will advance personalized care in patients with ALS.

Keywords

Amyotrophic lateral sclerosis, prediction, personalized care, gastrostomy

Submission date: 15 February 2024; Acceptance date: 22 May 2024

[1]Seoul National University College of Medicine, Seoul, Republic of Korea
[2]Department of Neurology, Neuroscience Research Institute, Medical Research Council, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Republic of Korea
[3]Department of Neurology, Neuroscience Research Institute, Medical Research Council, Seoul National University College of Medicine, SNU Boramae Medical Center, Seoul, Republic of Korea

Hyeon-Ji Oh and Won-Joon Lee contributed equally to this work.

Corresponding author:
Yoon-Ho Hong, Department of Neurology, Neuroscience Research Institute, Medical Research Council, Seoul National University College of Medicine, SNU Boramae Medical Center, 20 Boramaero-5-Gil, Dongjak-Gu, Seoul 07061, Republic of Korea.
Email: nrhong@gmail.com

Introduction

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease, characterized by progressive degeneration of motor neurons leading to muscle weakness, atrophy, spasticity and ultimately death typically in 3–5 years after symptom onset, unless respiratory support is provided.[1,2] One of the most common and debilitating symptoms in ALS is swallowing difficulty, with a reported cumulative incidence of up to 92%.[3] It can lead to potentially life-threatening complications such as dehydration, malnutrition, asphyxia and aspiration pneumonia.[3,4] Thus, proper evaluation of swallowing function is crucial in ALS care for preventing complications, maintaining nutrition and ultimately improving the patient's quality of life. For patients with severe dysphagia, intervention by gastrostomy tube insertion has been shown to significantly prolong survival and improve quality of life.[5–7]

Timely identification of the need for gastrostomy is crucial in advanced care planning, facilitating patient preparedness for forthcoming transitions and the preservation of autonomy for the maximal feasible duration. Nonetheless, determining the optimal timing for gastrostomy is a challenging process which involves an integrated evaluation of physical determinants such as nutritional status, and respiratory and swallowing functions, in addition to the social and psychological impacts of the intervention.[8,9] This challenge is further compounded by the broad phenotypic heterogeneity and variable trajectories of disease progression among patients. To date, research efforts have been largely focused on identifying prognostic factors at the group level, an approach that may not afford precise predictions for individual patients.[10,11]

Machine learning (ML) approach is particularly adept at building predictive models capable of navigating the complex and nonlinear relationships among variables—a task where traditional statistical methods may falter. These models also inherently possess capability to enhance their predictive accuracy iteratively through continuous training with new data. While a previous study has demonstrated the potential of ML for individualized survival prediction in ALS,[12] there is still a significant gap in exploratory efforts to employ ML for predicting specific clinical milestones including the loss of autonomy in swallowing function.

In this study, we aimed to develop an individualized prediction model to estimate the time to the loss of autonomy in swallowing function, leveraging the comprehensive large dataset from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database.[13] Further, we integrated the predictive models into a web application designed to estimate the critical milestone to facilitate a personalized and informed decision-making for gastrostomy intervention in ALS patients.

Methods

The study workflow is outlined in Figure 1. We developed individualized prediction models to estimate the time to reach the loss of autonomy in swallowing function in ALS patients, employing three different time-to-event algorithms: accelerated failure time (AFT), cox proportional hazards (COX) and random survival forest (RSF). For training and internal validation of the models, the PRO-ACT database was used. The target variable was defined as the time to a decline in the revised ALS Functional Rating Scale (ALSFRS-R) swallowing item score to 1 or below, indicating a need for supplementary tube feeding. The ALSFRS-R is a standard instrument for evaluating the functional status of patients with ALS, consisting of 12 items (speech, salivation, swallowing, handwriting, eating motion, dressing and hygiene, turning in bed and adjusting bed clothes, walking, climbing stairs dyspnea, orthopnea and respiratory insufficiency) in 4 domains (bulbar, gross motor, fine motor and respiratory) with each item scored between 4 (not affected) and 0 (worst).[14] Within the initial 3 months post-enrollment, predictive features were extracted and refined using embedded feature selection methods for each algorithm: lasso regularization in AFT and COX and permutation-based feature importance in RSF. Subsequently, the models were externally validated with data sourced from a tertiary referral center (Seoul National University Hospital, SNUH ALS/MND Registry), providing a real-world assessment of their predictive accuracy.

Study population

Data for training predictive models were sourced from the PRO-ACT database. The data available in the PRO-ACT Database have been volunteered by PRO-ACT Consortium members. PRO-ACT Dataset is the world's largest ALS clinical trial data repository, compiling placebo and treatment-arm data from 29 phase II/III clinical trials and 11,675 fully anonymized longitudinal Subject records funded by The ALS Therapy Alliance, Prize4Life, Inc., Northeast ALS Consortium (NEALS), Neurological Clinical Research Institute of Mass. General Hospital, ALS Finding A Cure and The ALS Association. Neurological Clinical Research Institute of Mass. General Hospital created and maintained the PRO-ACT Dataset and serves as the coordinating center and data distributor of the PRO-ACT Dataset. The database comprehensively features a range of clinical measurements across 17 categories, including adverse events, ALSFRS, concomitant medications, demographics, El Escorial criteria, family history, forced vital capacity (FVC) and others. We accessed the PRO-ACT database, which was last updated in August 2022, ensuring the use of the most current data available (https://nctu.partners.org/proact).
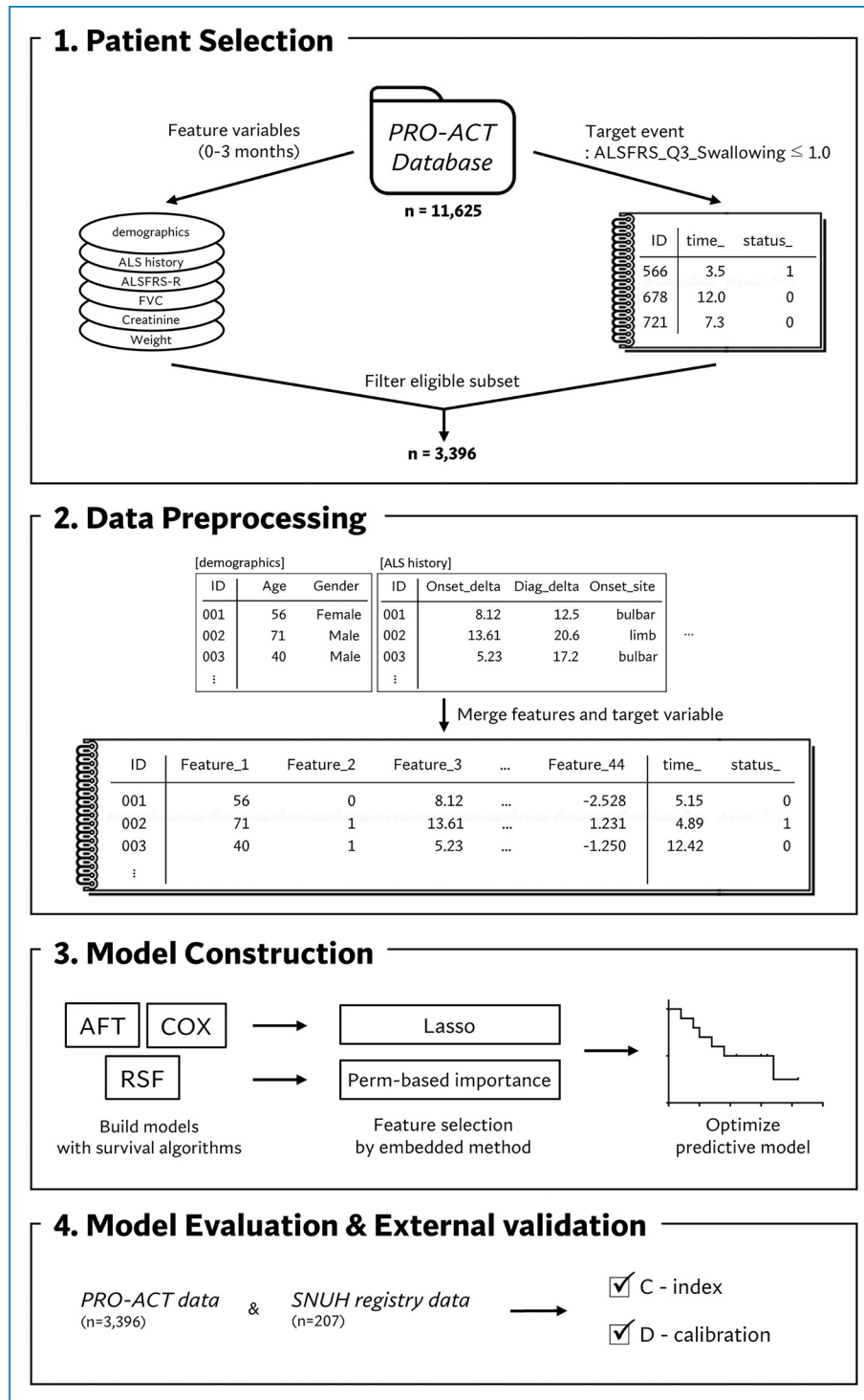
Figure 1. Overview of study workflow. (1) Patient selection. Within the PRO-ACT database, a subset of 3396 patients were selected, based on the presence of relevant predictive features and target variable data. (2) Data preprocessing. Both static (e.g. age, gender) and time-varying features (e.g. ALSFRS-R scores, weight, FVC) were extracted along with target variable data, and integrated into a comprehensive dataset where missing values were imputed using the Bayesian Ridge estimator, a multivariate imputation method. (3) Model construction. Three distinct predictive models were developed, each utilizing a unique algorithmic framework: AFT, COX and RSF. To optimize the predictive capability of the models, we employed embedded feature selection methods allowing for a tailored refinement of predictive features for each algorithm. (4) Evaluation of model performance and external validation. The predictive models underwent internal validation through the assessment of the concordance index (C-index) and distributional calibration (D-calibration) measures. Their generalizability was subsequently evaluated using an external dataset from a tertiary center ALS clinic registry.

We selected the records of 3396 patients from the database, according to the criteria based on the completeness of each participant's records for demographics, ALS history and ALSFRS-R. We excluded participants with incomplete or erroneous data entries, and the cases where the target event occurred either prior to or within the first 3 months following enrollment. For external validation, we utilized data from the SNUH ALS/MND registry, which included the patients registered from April 2017 through September 2022. The diagnosis of ALS was established based on the revised El Escorial criteria.[15] The inclusion criteria for our analysis were as follows: patients were required to have a diagnostic certainty level of at least clinically probable laboratory supported ALS, and three or more complete ALSFRS-R records with at least one of these records obtained a minimum of 3 months post their enrollment in the registry. Excluded were those patients who got gastrostomy or reached the target event either prior to or within the first 3 months following their registry enrollment.

Feature extraction and preprocessing

From the selected patient records, we extracted a set of features potentially related to the target variable. These included demographics (age, gender), ALS history (onset region, time from symptom onset, time since diagnosis), all ALSFRS-R item scores, FVC, blood creatinine level and body weight. Age was categorized into 5-year intervals, and the onset region was dichotomized into bulbar and nonbulbar categories. Diagnostic delay, defined as the time elapsed from symptom onset to diagnosis, was introduced as a meta-feature. For time-varying features such as ALSFRS-R, FVC, blood creatinine level and body weight, we calculated mean values and estimated slope values over the initial 3-month period. The slope values, representing the rate of change during this time frame, were determined using linear regression. To ensure the reliability of the slope estimation, we required a minimum of two data points spaced at least 1.5 months apart. We excluded the features with more than 40% missing data across patient records. To address remaining missing values, we employed a multivariate imputation method with the Bayesian Ridge estimator which models the target variable's probability distribution during training, effectively handling scenarios with sparse data.[16]

Model construction

We developed three distinct predictive models, each employing a different algorithmic approach: AFT, COX and RSF models. The AFT model, chosen for its parametric nature, assumes that the effect of covariates is multiplicative on the time scale of an event.[17] In the AFT framework, the Weibull distribution was used to model the relationship between covariates and the log of event time, as it can accommodate various hazard function shapes over time.

Unlike AFT, the COX model is a semi-parametric approach that does not assume a specific parametric form for the event time, but instead estimates how covariates impact the hazard ratio which is assumed to be constant over time.[18] The RSF is an extension of the Random Forest algorithm for survival analysis, combining predictions from multiple decision trees constructed using a random subset of features taking into account censored observations.[19] As a nonparametric method, it does not require any assumptions about the form of the hazard function or the distribution of the event time, thus providing even greater flexibility in modeling complex nonlinear relationships between covariates and the event outcome, although it is generally less interpretable and computationally intensive especially with large datasets.

The initial stage in constructing our predictive models involved hyperparameter optimization using grid search.[20] This method systematically explores a range of parameter combinations to identify the most effective setting for each model (Supplemental Information for Details). Subsequently, we refined feature selection using embedded methods which are particularly advantageous as they integrate feature selection seamlessly into the model training process.[21] This approach is instrumental in managing overfitting and computational complexity. For the AFT and COX models, we utilized the Least Absolute Shrinkage and Selection Operator (LASSO) technique, which minimizes a cost function combining the least squares error with an L1 norm penalty term ($\lambda\|\beta\|_1$) on the regression coefficients (β).[22] The LASSO cost function can be formulated as:

$$\sum_{i=1}^{N}\left(Y_i - \sum_{j=1}^{p} X_{ij}\,\beta_j\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right|$$

where N is the number of datapoints, p is the number of features and λ is a tuning parameter controlling the strength of the regularization. LASSO selectively shrinks certain coefficients towards zero, effectively blending feature selection with penalized regression. In the RSF model, feature selection was conducted using a permutation-based feature importance method. This method evaluates the significance of each feature on the model's predictive accuracy by measuring changes in accuracy when each feature's values are permuted.[23] The concordance index (C-index) was employed as the scoring function to assess permutation importance in the RSF model.

Evaluation of model performance

We evaluated the performance of predictive models in terms of both discrimination and calibration. Discriminative performance was quantified using the Harrell's concordance index (C-index) which is an extension of the area under the receiver operating characteristics (ROC) curve (AUC) adapted for

censored data, and evaluates the model's performance in accurately ranking pairs of subjects by their predicted risk.[24] A higher C-index signifies a model's enhanced ability to distinguish subjects who experience the event earlier from those who experience it later or not. The C-index is 0.5 for random guessing which can be considered as an implicit reference. For internal validation, a fivefold cross-validation (CV) procedure, repeated 10 times, was utilized to estimate the C-index for the predictive model using a dataset not involved in the model's training. To mitigate overestimation of model performance, missing data imputation was independently performed for each training and test set in every fold of the CV process. In external validation, bootstrap resampling was conducted with 1000 iterations and a sample size comprising 80% of the provided data. We also measured time-dependent ROC curves at different time points (6-month, 12-month and 24-month) to provide a more comprehensive picture of the models' discrimination performance across different follow-up times.[25]

To evaluate the congruence between observed outcomes and the predictions of our models, we assessed calibration performance using the Distributional Calibration method (D-Calibration), which, unlike traditional calibration methods focusing on specific time points, evaluates the entire distribution of predicted event times.[26] This method assesses the accuracy of predicted probabilities by examining whether a half of the subjects experience the event before the predicted median probability time and the other half afterward. D-Calibration extends this principle to multiple quantiles of the prediction curve, positing that the distribution of observed events should be uniform across these quantiles. In this method, predicted probabilities are segmented into 10 equal-sized groups or deciles. In an ideally calibrated model, each decile should represent an equal proportion of actual events, implying that 10% of events would be expected in each predicted probability decile.

Patient stratification

Patients from the PRO-ACT dataset were stratified into three progressor groups according to their predicted median event probability times: fast (first quartile), intermediate (second and third quartiles) slow (fourth quartile) progression group. For subjects without available median event probability time, linear extrapolation was employed to extend the event probability curve. For each stratification group, we calculated the average of the individual prediction curves generated by the models. These averages were then juxtaposed with the corresponding Kaplan–Meier (KM) survival curves for the same groups. The closeness between these two curves is indicative of the model's calibration accuracy.

Web application

We have developed an interactive web application, specifically designed to facilitate easy access to our predictive models by clinicians and researchers. This application primarily features the RSF model as its default option, while also providing users with flexibility to choose either AFT or COX model, depending on their specific requirements or preferences. A key functionality of the application is its ability to generate individualized predictions. Moreover, it displays average prediction curves categorized into three distinct ALS progression rate groups. These groups are classified based on the time required to reach a 50% event probability. Specifically: fast progression is assigned to the top 25th percentile, intermediate progression to the 25th–75th percentiles and slow progression to the bottom 25th percentile. The web application can be accessed at the following URL: https://predict-loss-of-autonomy-swallowing.streamlit.app/

Statistical analysis

Continuous variables are presented as means with standard deviations (or medians with interquartile ranges), and categorical variables were reported as counts (percentages). Differences between PRO-ACT and SNUH datasets were tested using the Mann–Whitney U test and Chi-squared test, respectively. Comparisons of C-index from three different models were made using ANOVA, followed by the pairwise t-test with Bonferroni correction for multiple group comparisons. For D-calibration, predicted event-free probability was segmented into 10 decile bins, with their uniformity across the entire dataset was analyzed using chi-square statistics. Pearson's chi-square test was employed to determine the uniformity of these bin. The efficacy of patient stratification by predictive models was evaluated through pairwise log-rank tests, comparing the KM curves of the respective stratified groups. For all tests, p-values were two-sided, and the significance level was set at 0.05.

Results

Study population

This study employed two distinct datasets: the PRO-ACT database for model development and internal validation, and the SNUH ALS/MND registry for external validation. Out of 11,625 de-identified patient records in the PRO-ACT database, we selected 3396 patients who had complete records in demographics, ALS history and ALSFRS-R scores. Similarly, from 394 ALS patients registered in the SNUH database, 207 patients were selected based on the same criteria for data completeness. The selection processes are detailed in the Supplemental Figure 1.

Table 1 provides a detailed comparison of baseline characteristics between the two study populations. Statistically significant differences were noted across a range of demographic and clinical features, suggesting the inherent

Table 1. Baseline characteristics of ALS patients in the PRO-ACT and SNUH datasets.

	PRO-ACT (n = 3396)	SNUH (n = 207)	*p*-values
Age (at enrollment)	57.0 (48.0–64.0)	60.0 (54.0–68.0)	<0.05
Gender (M:F)	2190:1206 (64.5%)	115:92 (55.6%)	0.08
Onset region			
Bulbar	612 (18.0%)	62 (30.0%)	<0.05
Nonbulbar	2784 (82.0%)	145 (70.0%)	
Time from symptom onset (months)	17.0 (11.5–24.2)	14.0 (9.0–21.5)	<0.05
Time from diagnosis (months)	5.6 (2.7–10.8)	2.0 (1.0–6.5)	<0.05
ALSFRS-R scores			
Total	38.0 (34.0–41.5)	40.0 (36.0–43.0)	<0.05
Bulbar	7.3 (6.0–8.0)	7.0 (6.0–8.0)	<0.05
Motor	16.0 (12.7–19.0)	18.0 (16.0–20.0)	<0.05
Respiratory	12.0 (11.0–12.0)	12.0 (11.0–12.0)	<0.05
Slope of ALSFRS-R (total, points per month)	−0.6 (−1.5–0.0)	−0.7 (−1.5–0.0)	0.86
Forced vital capacity (%)	83.3 (73.1–95.3)	80.5 (69.2–91.8)	<0.05
Weight (kg)	77.3 (67.0–88.6)	60.0 (54.0–67.0)	<0.05
Creatinine (serum, μmol/L)	65.4 (53.1–78.1)	63.9 (54.8–73.2)	0.25
Follow-up duration (months)	7.8 (3.0–10.2)	11.4 (4.9–19.4)	<0.05
Time to event (months)	6.0 (3.2–9.4)	11.4 (4.7–17.5)	<0.05
Event occurrence	615 (18.1%)	41 (19.8%)	0.94

Note. Summary statistics are presented as medians (interquartile ranges) for all continuous variables, and counts (percentages) for categorical variables. The values for time-resolved variables—including ALSFRS-R total, bulbar, motor, respiratory scores, FVC, weight, serum creatinine—are represented by their mean values calculated over the initial three months following enrollment. The slope of the ALSFRS-R total scores was estimated by fitting linear regression to the data collected during the first 3 months post-enrollment. The subdimension scores of the ALSFRS-R are determined as follows: the bulbar score is derived from the sum of items Q1 (speech) and Q3 (swallowing), excluding Q2 (salivation) due to its susceptibility to medication effects; the motor score comprises the aggregate of items Q4 (handwriting) through Q9 (climbing stairs); and the respiratory function score encompasses items R1 through R3.

disparities between data from controlled clinical trial and real-world registry settings. The onset region differed significantly, with a higher proportion of bulbar onset in the SNUH cohort compared to PRO-ACT (30.0% vs. 18.0%, *p* < 0.05). The intervals from symptom onset and from diagnosis to study entry were both shorter for the SNUH cohort, indicating a more rapid enrollment after symptom and disease identification. ALSFRS-R scores, which provide a granular assessment of patients' functional status, were higher in the SNUH group except for the bulbar domain subscore which was significantly lower compared to PRO-ACT (*p* < 0.05 for each). The baseline body weight was significantly lower in SNUH cohort compared to PRO-ACT (*p* < 0.05). The median follow-up duration in the PRO-ACT dataset was notably shorter at 7.8 months (IQR, 3.0–10.2 months), compared to the 11.4 months in the SNUH cohort (IQR, 4.9–19.4 months). Analysis of the KM curves revealed a statistically significant delay in the occurrence of the target event for the SNUH cohort relative to that observed in the PRO-ACT (*p* < 0.05 with log-rank test), as illustrated in Supplemental Figure 2.

Predictive feature selection

Through the model optimization process, we selected distinct sets of predictive features, aligning them with each algorithm to enhance the predictive accuracy and robustness of our models. Notably, a significant convergence in the selection of key variables was observed across all three algorithms, including age, time from symptom onset, mean ALSFRS-R scores for salivation and swallowing, mean bulbar subscale score and mean FVC. This consistency, particularly in the inclusion of bulbar-related features and the exclusion of lower extremity-related items in the feature selection process, highlights the significant role of bulbar dysfunction in predicting dysphagia progression in ALS patients. The results of feature selection and the predictive feature contribution for each model are illustrated in Supplemental Figures 3–7.

Model performance

The performance of predictive models was evaluated in terms of two key metrics: discrimination—the ability to differentiate patients who reached the endpoint earlier versus those who did later or not at all, and calibration which refers to the agreement between observed and predicted times to the endpoint. Figure 2 depicts the analysis of discrimination as measured with C-index and time-dependent ROC curves (Figure 2(a) and (c)). In the internal validation using PRO-ACT data, all three models exhibited strong discriminative capability with no significant difference between models: the median (IQR) C-index of 0.851 (0.842–0.859) for AFT, 0.850 (0.841–0.859) for COX and 0.846 (0.839–0.854) for RSF. In the external validation with bootstrapped data from the SNUH cohort, these models consistently demonstrated effective discrimination, but with statistically significant difference between models ($p < 0.05$), as evidenced by the median (IQR) C-index of 0.803 (0.776–0.826), 0.802 (0.774–0.826) and 0.785 (0.755–0.811) for AFT, COX and RSF models, respectively. The results of time-dependent ROC analysis showed that the discrimination performance of the models tended to decrease over time (Figure 2(b) and (d)). In the PRO-ACT data, RSF performance remained relatively stable, while the AUC values of COX and AFT decreased noticeably at 24 months. In contrast, in the SNUH dataset for external validation, the performance of all models decreased from month 12, with RSF in particular exhibiting a further decrease at 24 months, suggesting a potential overfitting of RSF especially in the later stages of disease progression.

Figure 3 demonstrates the results of this calibration analysis. In both the internal and external validations, all models exhibited effective distributional calibration, as evidenced by the uniform distribution of observed events across each decile. While not reaching statistical significance, a left-skewness pattern was observed in the external validation, suggesting a tendency of the models to overestimate event probabilities in this external cohort.

Patient stratification

To evaluate the stratification capacity of our models, patients in the PRO-ACT dataset were classified into three groups (fast progression, intermediate progression and slow progression) based on the median predicted event probability. Comparative analysis of the KM curves for these groups revealed significant differences in all pairs of stratified groups for each model ($p < 0.05$) (Figure 4). Furthermore, to evaluate the concordance between the predicted and actual time-to-event data, we juxtaposed the mean prediction curves within each progression group with their respective KM curves. A marked proximity between these curves was observed over the course of time, especially in the case of the AFT and COX models, indicating a high degree of calibration accuracy in these models. The RSF model, however, shows a distinct pattern where the KM curve for the slow progressor group did not show a precipitous drop over time. In addition, the RSF model's prediction curves of the fast and intermediate progressor groups showed a tendency towards a higher event-free probability over time compared to the corresponding KM curves.

Discussion

This study addresses the critical challenge of individualized prognosis for clinical milestones in a progressive neurodegenerative disease, focusing on the loss of autonomy in swallowing function in ALS. Leveraging eminent time-to-event prediction algorithms and the largest clinical trials database, we built three prediction models and confirmed their discrimination and calibration performance through external validation using the real-world data from a tertiary hospital registry. Furthermore, we integrated our models into a web application specifically designed to estimate the critical milestone, which could facilitate personalized care and inform clinical trial designs through individualized risk assessments for ALS patients.

We strategically employed various algorithms, each paired with a distinct set of predictor variables, for the development of predictive models. This approach facilitated a comprehensive exploration of a range of analytical perspectives, enabling us to uncover unique insights that a single algorithm or a uniform set of predictors might overlook. Notably, across all three algorithms, there was a significant convergence in the selection of key predictor variables. This convergence, especially evident in the consistent emphasis on bulbar-related features and the exclusion of lower extremity-related items across different models, highlights the crucial role of bulbar dysfunction in the progression of dysphagia and illuminates the regional pattern of
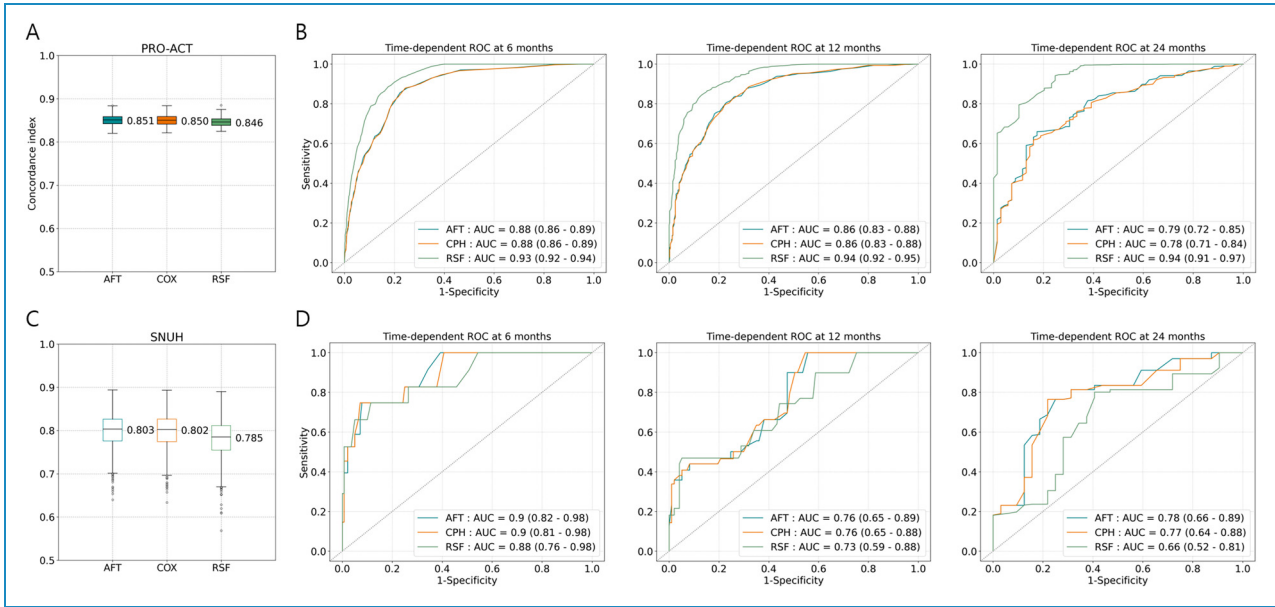
Figure 2. Discriminative performance of the AFT, COX and RSF models as measured by using concordance index (C-index) (a and c) and ROC curves (b and d). The results of internal validation indicated no significant difference between the three models (a), whereas external validation demonstrated significant variability in C-index values across the three models ($p < 0.05$) (c). Shown are the ROC curves of the models at different time points (6-month, 12-month and 24-month) in the PRO-ACT (b) and SNUH cohort (d). The area under the ROC curve (AUC) is presented along with 95% confidence intervals (CIs), which were estimated using bootstrap resampling performed 1000 times.
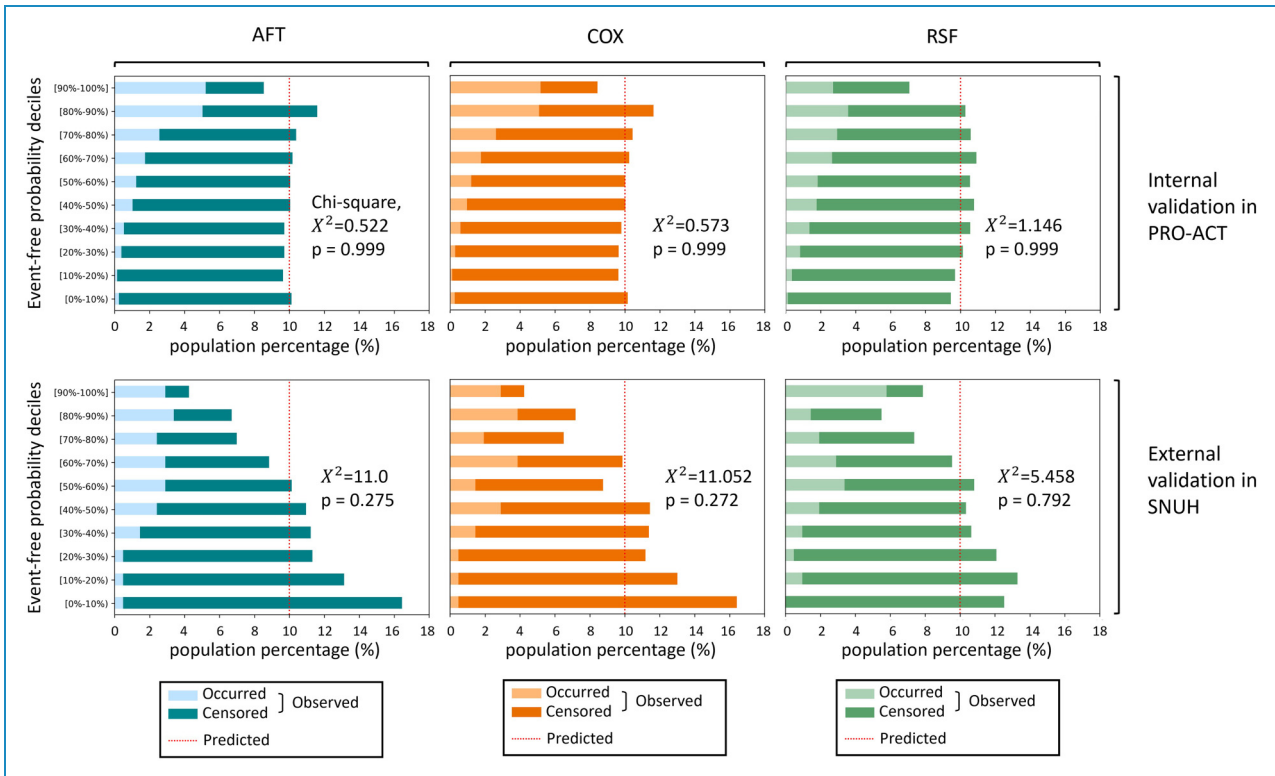
Figure 3. Distributional calibration histograms for the AFT, COX and RSF models. Event-free probabilities are binned into 10 deciles, and each horizontal bar in the histogram represents the observed event rate in each respective decile bin. For an ideally calibrated model, these bars would consistently align with a 10% event rate. The chi-square test was utilized to determine the statistical significance of the calibration across the deciles. The models were validated internally using the PRO-ACT dataset and externally with the SNUH dataset.
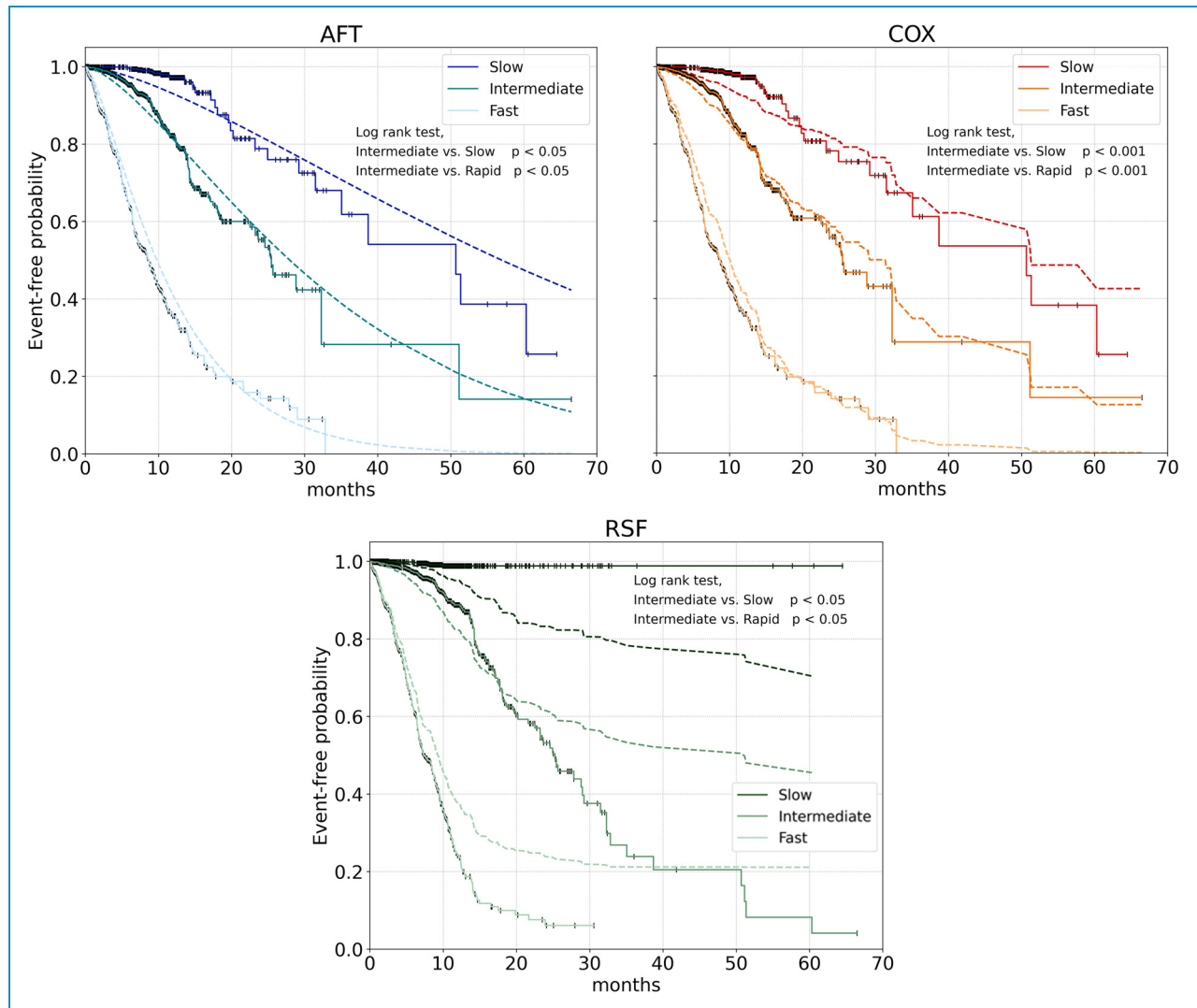
Figure 4. Patient stratification by predictive models. Cases from the PRO-ACT dataset were stratified into three progressor groups (fast, intermediate and slow) based on the median event probability times predicted by the models. Solid lines represent the KM curves for each progressor group, and dotted lines depict the mean prediction curves within the corresponding groups. The KM curves are evaluated across stratified groups with pairwise log-rank tests.

disease progression in ALS patients. On the other hand, we also found differences in the results of feature selection, especially between RSF and the other models. Despite being highly correlated with swallowing difficulties, the speech disability was selected as an important feature only in the RSF model. This may be related to the complexity of data structure and the differences between algorithms in how they handle the data. Penalization in AFT and COX models can shrink coefficients of highly correlated features to zero, effectively removing one of them from the model due to redundancy. In contrast, RSF can retain both correlated features due to their combined predictive power. In further analysis of predictive models, we observed that two key features—the mean values of ALSFRS-R swallowing item score and FVC in the first 3 months—contravened the

proportional hazards assumption, indicating the importance of examining the specific assumptions of the algorithm in the process of feature selection.

While all models were effective at stratifying patients, the results of RSF model, unlike AFT and COX, did not match well with the KM curves of the corresponding risk groups. The RSF's nonparametric, ensemble-based methodology is generally believed to capture effectively complex nonlinear relationships between variables, including uncertainty due to censored observations.[19] These properties might have resulted in more conservative event probability estimates compared to KM curves for the fast and intermediate progressor groups. On the other hand, as for the misalignment in slow progressors, it could be explained by the fact that the vast majority (99%) of

those classified as slow progressors by the RSF model were censored subjects. Since the KM estimator treats censored data as having no event, it may account for the flat KM curve of the slow progressor group.[27] Taken together, our results suggest that it is important to be mindful of potential biases in the context of censored data with complex structure. These insights collectively underscore the importance of selecting appropriate algorithm based on the specific characteristics of the dataset, providing invaluable guidance for future research. By understanding these nuances and cross-referencing the results of different models, we believe that the multialgorithm approach could provide more comprehensive insights into the accuracy and uncertainty of predictions.

This study follows the path of previous research in developing individualized prediction models in ALS,[12,28,29] but the key difference is that the prediction is targeted at clinical milestone, specifically loss of autonomy in swallowing function. Not only is this highly relevant to the selection of predictive features, but also will have a practical impact on the decision-making process regarding gastrostomy in ALS patients. While there was no significant difference between the time to the loss of autonomy in swallowing and actual gastrostomy placement, we noticed a discernible delay in gastrostomy intervention for the SNUH cohort (Supplemental Figure 8). This suggests the possibility of additional factors influencing the decision-making process which requires careful consideration of not only physical impact but also social and psychological burden.[30] While gastrostomy can significantly prolong survival and improve quality of life by addressing nutritional challenges,[6] it also involves a compromise in bodily autonomy. The social and psychological implications of using and maintaining a feeding tube, which is often perceived as a "symbol of deterioration" akin to wheelchairs or ventilators, are profound.[31] Therefore, it is imperative to engage in a detailed and nuanced discussion to ascertain the most beneficial timing for gastrostomy, striking a balance between its advantages and the preservation of bodily independence. In light of these considerations, we propose the application of our predictive models as a useful tool to support this critical decision-making process for gastrostomy in ALS patients.

Our online models should be applied with an understanding of their potential constraints, especially in terms of dataset representativeness and algorithmic variations. First, our models were primarily trained using data from the PRO-ACT database, which, despite being the most extensive compilation from clinical trials, may not wholly represent the general ALS population due to specific eligibility criteria inherent in clinical trials. Consequently, the model's applicability to the broader ALS population could be restricted.[32] The external validation using the SNUH dataset highlighted a notable limitation: all models exhibited a left-skewed distribution in the decile histogram of D-calibration, suggesting an overestimation of event risk. This skewness may be attributable to the distinct

characteristics of the clinic-based population, including the shorter interval from symptom onset and diagnosis coupled with a better baseline functional status, which could influence the model outcomes. Second, employing a variety of algorithms and predictor sets introduces complexity in interpreting results and comparing models. There are pros and cons to using the same feature selection process for different algorithms.[33] Although using the same feature selection process makes it more straightforward to compare different models, we might miss potentially relevant features for a particular algorithm. Since the main goal of feature selection in this study was to improve the overall performance of each predictive model, we made the feature selection process an integral part of the model optimization and chose a popular method for each algorithm. Acknowledging that the results do not provide algorithmic benchmarks with the same feature set, we believe that they provide valuable insights into the performance capabilities of each model when optimized individually. Lastly, it should be recognized that our models were trained with the exclusion of cases where the target event occurred within the first three months of enrollment, so they may be of limited use in patients who already have significant swallowing difficulties.

In conclusion, we developed individualized prediction models for the loss of autonomy in swallowing function in ALS patients. These models, leveraging three distinct algorithms, demonstrated robust discriminatory and calibration capabilities in both internal evaluation and external validation, underscoring their reliability and applicability. Our models are implemented into a web application which could be applied to individualized counselling, management and clinical trial design for gastrostomy intervention. While acknowledging the inherent limitations concerning the representativeness of the training data and the complexities introduced by multiple algorithms and diverse predictor sets, we believe that our models can aid clinicians in making more informed decisions, by offering individualized risk assessments in the context of gastrostomy intervention for ALS patients. We anticipate that further studies will focus on enhancing the generalizability of these models and exploring their application in diverse clinical settings to facilitate personalized care in ALS patients.

ORCID iD: Yoon-Ho Hong https://orcid.org/0000-0002-3325-6358

References

1. Feldman EL, Goutman SA, Petri S, et al. Amyotrophic lateral sclerosis. *Lancet* 2022; 400: 1363–1380.
2. Spittel S, Maier A, Kettemann D, et al. Non-invasive and tracheostomy invasive ventilation in amyotrophic lateral sclerosis: utilization and survival rates in a cohort study over 12 years in Germany. *Eur J Neurol* 2021; 28: 1160–1171.
3. Perry BJ, Nelson J, Wong JB, et al. The cumulative incidence of dysphagia and dysphagia-free survival in persons diagnosed with amyotrophic lateral sclerosis. *Muscle Nerve* 2021; 64: 83–86.
4. Desport JC, Preux PM, Truong TC, et al. Nutritional status is a prognostic factor for survival in ALS patients. *Neurology* 1999; 53: 1059–1063.
5. Burkhardt C, Neuwirth C, Sommacal A, et al. Is survival improved by the use of NIV and PEG in amyotrophic lateral sclerosis (ALS)? A post-mortem study of 80 ALS patients. *PLoS One* 2017; 12: e0177555.
6. Fasano A, Fini N, Ferraro D, et al. Percutaneous endoscopic gastrostomy, body weight loss and survival in amyotrophic lateral sclerosis: a population-based registry study. *Amyotroph Lateral Scler Frontotemporal Degener* 2017; 18: 233–242.
7. Mazzini L, Corrà T, Zaccala M, et al. Percutaneous endoscopic gastrostomy and enteral nutrition in amyotrophic lateral sclerosis. *J Neurol* 1995; 242: 695–703.
8. Andersen PM, Borasio GD, Dengler R, et al. EFNS Task force on diagnosis and management of amyotrophic lateral sclerosis: guidelines for diagnosing and clinical care of patients and relatives. *Eur J Neurol* 2005; 12: 921–959.
9. Miller RG, Jackson CE, Kasarskis EJ, et al. Practice parameter update: the care of the patient with amyotrophic lateral sclerosis: drug, nutritional, and respiratory therapies (an evidence-based review) report of the quality standards subcommittee of the American academy of neurology. *Neurology* 2009; 73: 1218–1244.
10. Kjældgaard AL, Pilely K, Olsen KS, et al. Prediction of survival in amyotrophic lateral sclerosis: a nationwide, danish cohort study. *BMC Neurol* 2021; 21: 1–8.
11. Moura MC, Novaes MRCG, Eduardo EJ, et al. Prognostic factors in amyotrophic lateral sclerosis: a population-based study. *PLoS One* 2015; 10: e0141500.
12. Westeneng H-J, Debray TPA, Visser AE, et al. Prognosis for patients with amyotrophic lateral sclerosis. *Lancet Neurol* 2018; 17: 423–433.
13. Atassi N, Berry J, Shui A, et al. The PRO-ACT database: design, initial analyses, and predictive features. *Neurology* 2014; 83: 1719–1725.
14. Franchignoni F, Mora G, Giordano A, et al. Evidence of multidimensionality in the ALSFRS-R scale: a critical appraisal on its measurement properties using Rasch analysis. *J Neurol Neurosurg Psychiatry* 2013; 84: 1340–1345.
15. Brooks BR, Miller RG, Swash M, et al. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Other Motor Neuron Disord* 2000; 1: 293–299.

16. Tipping ME. Sparse Bayesian learning and the relevance vector machine. *J Mach Learn Res* 2001; 1: 211–244.

17. Swindell WR. Accelerated failure time models provide a useful statistical framework for aging research. *Exp Gerontol* 2009; 44: 190–200.

18. Orbe J, Ferreira E and Núñez-Antón V. Comparing proportional hazards and accelerated failure time models for survival analysis. *Stat Med* 2002; 21: 3493–3510.

19. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *Ann Appl Stat* 2008; 2: 841–860.

20. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–2830.

21. Saeys Y, Inza I and Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; 23: 2507–2517.

22. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Series B* 2011; 73: 273–282.

23. Strobl C, Boulesteix AL, Kneib T, et al. Conditional variable importance for random forests. *BMC Bioinformatics* 2008; 9: 307.

24. Pencina MJ and D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004; 23: 2109–2123.

25. Kamarudin AN, Cox T and Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med Res Methodol* 2017; 17: 53.

26. Andres A, Montano-Loza A, Greiner R, et al. A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PLoS One* 2018; 13: e0193523.

27. Leung KM, Elashoff RM and Afifi AA. Censoring issues in survival analysis. *Annu Rev Public Health* 1997; 18: 83–104.

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

29. Huang Z, Zhang H, Boss J, et al. Complete hazard ranking to analyze right-censored data: an ALS survival study. *PLoS Comput Biol* 2017; 13: e1005887.

30. Son B, Lee J, Ryu S, et al. Timing and impact of percutaneous endoscopic gastrostomy insertion in patients with amyotrophic lateral sclerosis: a comprehensive analysis. *Sci Rep* 2024; 14: 7103.

31. White S, O'Cathain A, Halliday V, et al. Factors influencing decisions people with motor neuron disease make about gastrostomy placement and ventilation: a qualitative evidence synthesis. *Health Expect* 2023; 26: 1418–1435.

32. Chiò A, Canosa A, Gallo S, et al. Do enrolled patients accurately represent the ALS population? *ALS Clin Trials* 2011; 77: 1432–1437.

33. Pudjihartono N, Fadason T, Kempa-Liehr AW, et al. A review of feature selection methods for machine learning-based disease risk prediction. *Front Bioinform* 2022; 2: 927312.