CHALLENGE LAB
CLINICAL DATA ANALYSIS IN ALS PATIENTS

OBJECTIVE

The objective of this project is to analyze clinical data from ALS (Amyotrophic Lateral Sclerosis) patients in order to model disease severity and progression. The project involves both methodological and practical challenges related to modeling longitudinal medical data and selecting appropriate response variables and statistical frameworks.

PROJECT TRACKS

1. DATA UNDERSTANDING AND PREPROCESSING

Objective: Gain familiarity with the ALS dataset collected at the ALS Center of Federico II University Hospital of Naples, led by Prof. Raffaele Dubbioso.

Dataset Description: The ALS dataset includes 254 patients (154 male, 100 female), with a total of 1,412 visits. Each patient underwent multiple clinical evaluations (ranging from 1 to 34 visits), which include:

- **Demographic and anamnesic variables:** sex, date of birth, height, weight, family history, age at onset.
- **Clinical variables:** ALSFRS-R, MRC scores, PUMNS, KINGS stage, respiratory measures (FVC, ventilation, tracheostomy), PEG status.
- **Disease progression:** duration, rate of progression, diagnostic delay, and type of therapy.

A detailed description of a subset of the variables is reported in Table 1.

Before proceeding with modeling, students are expected to conduct a thorough preliminary analysis of the dataset. This step is crucial to understand the data structure, assess data quality, and make informed decisions for subsequent analyses.

Handling Missing Data: The dataset contains missing values (NA). Students must decide whether to exclude these missing observations or to apply appropriate imputation methods.

Descriptive Statistics: Students should compute and report descriptive statistics to summarize the main characteristics of the data. This includes measures of central tendency

Variable	Description
Patient_ID	Unique identifier for each patient
Visit_ID	Unique identifier for each visit
Date_of_birth	Patient's date of birth
Date_symptom_onset	Date of symptom onset
Date_of_diagnosis	Official diagnosis date
Sex	Gender (0 = male, 1 = female)
Age_at_onset	Age at symptom onset
Clinical_onset	Initial clinical manifestation
Spirometry	Ability to perform test (0, 1, NA)
FVC	Forced vital capacity category
Ventilation	Use of mechanical ventilation (coded 0–3)
Tracheostomy	Presence of tracheostomy (0 = no, 1 = yes)
PEG	PEG status (0 = no, 1 = yes, 2 = refused but needed)
Family_history	Family history of ALS (0 = no, 1 = yes)
Therapy	ALS-specific therapy (riluzole, edaravone, both, none)
Diagnostic_delay	Delay between symptom onset and diagnosis (months)
MRC_UL	Upper limb muscle strength (0–70)
MRC_LL	Lower limb strength (0–60)
MRC_Bulbar	Bulbar region strength (0–15)
PUMNS_Bulbar	UMN burden (bulbar)
PUMNS_UL	UMN burden (upper limbs)
PUMNS_LL	UMN burden (lower limbs)
CNS_LS	CNS involvement score
KINGS_TOT	King's staging score
Disease_duration	Disease duration (months)
Rate_of_progression	(48 - ALSFRS_R) / Disease_duration
ALSFRS_R	Functional Rating Scale
ALSAQ_5	Quality of life questionnaire

Table 1: Subset of ALS dataset variables

(mean, median), dispersion (standard deviation, interquartile range), and distribution shape (skewness, kurtosis) for continuous variables. For categorical variables, frequency tables and proportions should be presented. Visualizations such as histograms, boxplots, and scatterplots are encouraged to identify data distributions, outliers, and potential relationships between variables. This exploratory data analysis will aid in understanding the data and in guiding the choice of appropriate statistical models.

Data Transformation: If variables exhibit skewed distributions or violate model assumptions, appropriate data transformations (e.g., logarithmic, square root) should be considered. Students should justify the use of transformations and assess their impact on the data distribution and subsequent analyses.

2. RESPONSE VARIABLE SELECTION

Students must choose one of the following response variables based on the research objective:

- **KINGS_TOT** – Clinical staging score. It is a categorical variable with an inherent order (0, 1, 2, 3, 4A, 4B), but students must decide whether to treat it as an *ordinal outcome* or as a *nominal/disconnected categorical outcome*. The choice must be clearly stated and statistically justified based on data distribution, model assumptions, and interpretability.
  Additionally, students may choose to **collapse categories**, provided that they *explicitly justify the decision* using clinical reasoning, exploratory analysis, or modeling considerations (e.g., low sample size in some categories).
- **Rate_of_progression** – Progression rate (continuous)
- **ALSFRS_R** – Functional rating scale (continuous, bounded between 0 and 48)
- **ALSAQ_5** – Quality of life score (continuous, self-reported outcome)

3. ANALYSIS DESIGN CHOICE

You must decide how to treat statistical units:

- **Cross-sectional approach:** All visits are treated as independent observations. That is, each row in the dataset (each patient-visit combination) is considered as a separate statistical unit, regardless of patient identity. This simplifies the analysis but ignores the correlation between repeated measures on the same patient.
- **Longitudinal approach:** The repeated visits per patient are modeled using appropriate statistical techniques to account for intra-subject correlation.

4. MODEL SELECTION AND COMPARISON

Based on your selected response variable and data structure, apply one or more of the following:

- Generalized Linear Models (GLM) appropriate
- Mixed-effects models for longitudinal data
- Model comparison (e.g., GLM vs. mixed model, comparison between different GLMs)

Include justification for each modeling choice. You may also include diagnostic plots, AIC/BIC comparisons.

DATA ACCESS

Due to privacy restrictions, the dataset is not public. Students who choose this case study must request access from the course instructor.

DEADLINE

The final version of the project must be submitted by the exam date.

GROUP WORK

This project can be completed individually or in groups. Collaboration is encouraged, but each member must clearly state their contribution in the final report.

REFERENCES

For mixed models applied to longitudinal data analysis, please refer to the book available in the “Books” folder on our Teams space: *Multilevel Modelling for Public Health and Health Services Research*. For background and recent developments on ALS disease, consult the scientific articles provided in the “ALS Case Study” folder. Additional sources may be used as long as they are properly cited.

DELIVERABLES

Each group must submit a scientific report structured as a research article. The report must include the following sections:

1. **Introduction**
   - Provide clinical and scientific context on ALS and the importance of understanding disease progression.
   - Clearly state the objective(s) of the study.
   - Include a brief review of relevant literature or previously used methods for modeling ALS progression or staging.
2. **Materials and Methods**
   - Describe the dataset, including sample size, variable types, and structure (e.g., longitudinal).
   - Explain the data preprocessing steps (e.g., treatment of missing data, variable selection, data transformations).
   - Justify the choice of the response variable and modeling approach (e.g., GLM, mixed models).
3. **Results & Discussion**
   - Present the main results, including performance metrics and model comparisons if multiple approaches are tested.

- Interpret the findings from a clinical and statistical perspective.
- Discuss limitations of the data or methodology, and potential sources of bias.

4. **Conclusion and Future Work**

    - Summarize key takeaways and insights.
    - Suggest possible extensions, improvements, or alternative approaches for future research.

5. **Contributions Declaration**

    Each team member must clearly declare their contribution to the project, specifying:

    - Who worked on data preprocessing and feature selection.
    - Who developed and tuned the models.
    - Who performed the evaluation and comparative analysis.
    - Who was responsible for writing and structuring the final report.

6. **References**

Note: The final submission must include:

- A scientific paper (PDF format, 6–10 pages).
- The complete codebase (R or Python).
- A slide deck (PDF or PowerPoint) summarizing the main points, results, and conclusions (max 15/20 slides) to present at the exam.