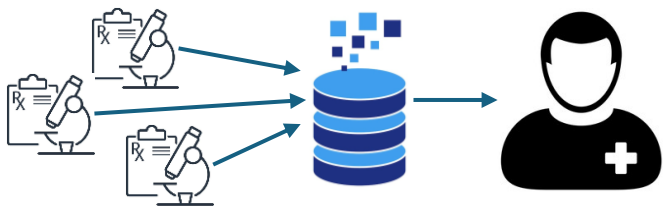
# PREDICTING ALS PROGRESSION WITH AI: INSIGHTS FROM CLINICAL AND BIOLOGICAL MARKERS

Lucas Bouclier, Christel Dartigues-Pallez, Johan Montagnat

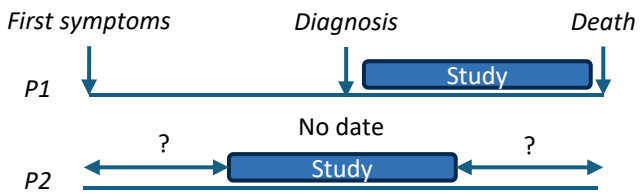
The aim of this work is to predict the amyotrophic lateral sclerosis (ALS) progression in patients from clinical and biological data using artificial intelligence algorithms. We estimate patients' health state evolution three months ahead and identify the medical features most correlated with disease progression.

**PROACT** includes 16 tables with clinical, biological, and physiological measurements, functional scores, treatments, demographics, and family history. It also includes the **ALSFRS-R score** (0–48), which is the reference for clinical follow-up. The dataset is **heterogeneous, sparse, and biased, with missing values** due to its 23 different sources.

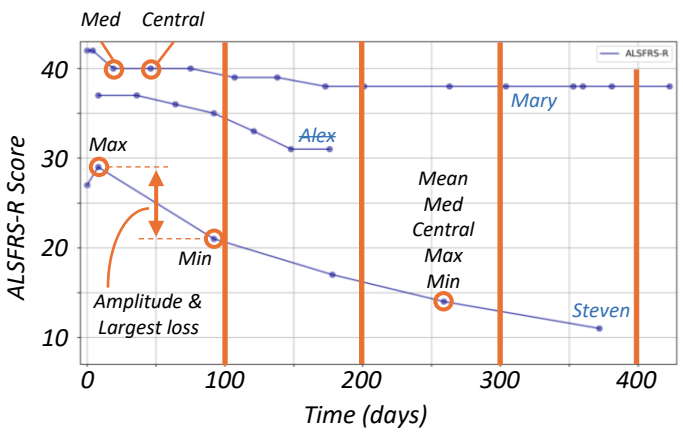


*23 international clinical trials* — ***PROACT*** *database* — *11675 patients*

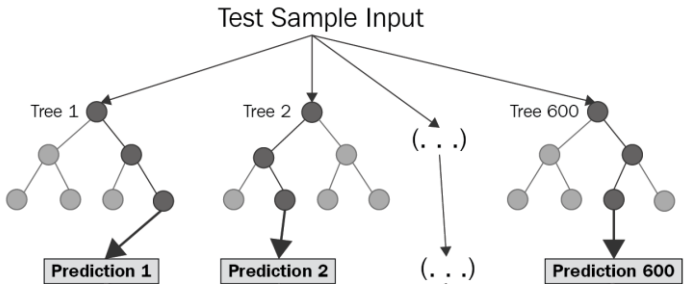
We also have a **temporal data alignment issue**. Here we compare two patients. P1 has dates for symptom onset, diagnosis, and death, while P2 does not.




We **sampled observations by quarter**, consistent with the quarterly monitoring of patients at the hospital. For each quarter, we extracted the max, min, mean, median, central, and amplitude values, as well as the largest loss between two observations. Patients without at least one observation per quarter were excluded.




We used **Random Forest** models, which offer **strong predictive performance** and allow us to **interpret the importance of each individual variable**, even with relatively small cohorts and hundreds of features.




We trained 3 successive models using the **first one, two, or three quarters to predict the next quarter's ALSFRS-R score**. Our results are compared to regression methods and a naive method subtracting the average quarterly decline from each patient's last observation.

| | $Q_1 \Rightarrow Q_2$ | $Q_{1,2} \Rightarrow Q_3$ | $Q_{1\text{-}3} \Rightarrow Q_4$ |
|---|---|---|---|
| Naive method | MAE : 2.49<br>RMSE : 3.28<br>$R^2$ : 0.78 | MAE : 2.66<br>RMSE : 3.53<br>$R^2$ : 0.82 | MAE : 2.84<br>RMSE : 3.82<br>$R^2$ : 0.84 |
| Linear regression | MAE : 3.13<br>RMSE : 4.56<br>$R^2$ : 0.58 | MAE : 2.39<br>RMSE : 3.35<br>$R^2$ : 0.84 | MAE : 2.59<br>RMSE : 3.56<br>$R^2$ : 0.86 |
| Sigmoid regression | MAE : 3.38<br>RMSE : 5.44<br>$R^2$ : 0.48 | MAE : 2.57<br>RMSE : 3.57<br>$R^2$ : 0.82 | MAE : 2.62<br>RMSE : 3.75<br>$R^2$ : 0.85 |
| Random Forest Score only | MAE : 2.31<br>RMSE : 3.13<br>$R^2$ : 0.80 | MAE : 2.28<br>RMSE : 3.14<br>$R^2$ : 0.86 | MAE : 2.58<br>RMSE : 3.47<br>$R^2$ : 0.87 |
| Random Forest All data | MAE : **2.11**<br>RMSE : **2.87**<br>$R^2$ : **0.83** | MAE : **2.13**<br>RMSE : **2.90**<br>$R^2$ : **0.88** | MAE : **2.47**<br>RMSE : **3.34**<br>$R^2$ : **0.88** |

Among the biological variables analyzed, biomarkers such as **creatinine, creatine kinase, and bicarbonate** appear to be the most informative for predicting disease progression. Ongoing work aims to reduce prediction error and build confidence in these predictors by improving data processing, selecting the most relevant clinical and biological data, and comparing performance with other machine learning methods.

Contact: lucas.bouclier@univ-cotedazur.fr

cnrs — UNIVERSITÉ CÔTE D'AZUR — i3S sophia antipolis