Inspecting Progression Trajectories in Amyotrophic Lateral Sclerosis using Process Mining

Erica Tavazzi[1*], Roberto Gatta[2], Mauro Vallati[3], Stefano Cotti Piccinelli[2], Massimiliano Filosto[2], Maurizio Castellano[2], Barbara Di Camillo[1,4]

[1] Department of Information Engineering, University of Padova, Padova, Italy. erica.tavazzi@unipd.it
[2] Department of Clinical and Experimental Sciences, University of Brescia, Brescia, Italy.
[3] School of Computing and Engineering, University of Huddersfield, Huddersfield, United Kingdom.
[4] Department of Comparative Biomedicine and Food Science, University of Padova, Padova, Italy. *corresponding author

Keywords: Amyotrophic lateral sclerosis, process mining, disease trajectories, patient stratification, prognostic biomarkers.

Abstract. Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease whose mechanisms are still fully unclear. Being able to predict ALS prognosis would help in improving the patients' quality of life and support clinicians in planning treatments. On the one hand, most of the modeling approaches to ALS miss to catch the evolving nature of the disease; on the other, Process Mining (PM) comprehends techniques useful to generally describe processes, but often misses methods to reveal statistically significant differences in the mined pathways. In this paper, we investigate ALS evolution using PM techniques enriched to easily mine processes and, at the same time, automatically reveal how the pathways differentiate according to patients' characteristics.

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) is a rare neurological disease that primarily affects motor neurons, causing progressive paralysis of most voluntary muscles and usually leading to death for respiratory insufficiency within 3-5 years from onset. In ALS, the pathogenic mechanisms as well as the rate of progression and impairment patterns are still unclear, making its diagnosis as well as the development of therapies or intervention plans very challenging [1].

In this context, developing data-driven tools able to model the progression of ALS could help to describe the manifold nature of this disease, to identify risk factors, to group patients based on similar evolution patterns, and can become a mean for personalized predictive purposes. For this reason, there is a growing interest in methods for mining and analyzing the progression of ALS, particularly by considering longitudinal clinical data: examples include Neural Networks, Sequential Pattern Mining, and Dynamic Bayesian Networks [2, 3, 4].

In this work, we consider a different perspective by adopting a process-oriented approach for mining, describing, and predicting the progression of ALS in a clinical trial population. First, we structure the patients' longitudinal information collected during the trial as an Event Log (EL). EL is a sequence of tuples <patient ID, Event, Datetime, *Attributes*>, where the Event can be any clinical or relevant event occurred in the patient's life (*e.g.*, medical examination, a new impairment, death) and *Attributes* is an optional set of features specific for that event (*e.g.*, a set of numeric values for an event of type 'Lab exam', or the grade of an adverse event for an event of type 'Drug

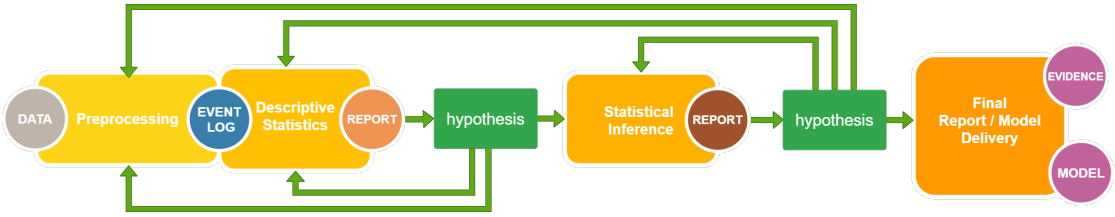

Figure 1: The computational pipeline exploited for the analysis.

administration'). Then, by employing Process Mining (PM) techniques such as the Direct Following Diagram (DFD) and the CareFlow Miner (CFM), we aim at discovering the underlying models that generated the patients' EL. We are then in the position to analyze the patterns of evolution of the patients, and investigate the predictive potential of the mined processes for describing and forecasting the prognosis – in terms of timing or sequence of events – based on the value of specific covariates at baseline.

2 Materials and Methods

The traditional PM computational pipeline can be summarized as follows:

- *Preprocessing*: the data are processed and shaped in the form of an EL. More ELs can be produced, at this point, to highlight different aspects of the data.
- *General Descriptive Statistics*: it is performed to identify volumes, general data distribution and reveal statistical biases. Here we used DFD and CFM to perform this task. The produced reports are commented with domain experts, and are used to formulate hypotheses.
- *Statistical Inference*: it is performed to identify and reveal statistical dependencies and correlations, with indicators such as p-values or confidence intervals, among data. This is a pivotal step of the pipeline, that allows to confirm or reject formulated hypotheses, and to support the formulation of further theses as well.
- *Final Report/Model delivery*: if evidences are supported by the help of domain experts, communicative reports can be produced; Decision Support Systems models can be made usable and delivered to the final users.

Figure 1 provides an overview of the described pipeline. Each descriptive or inferential step represents an opportunity to refine the analysis by formulating and testing new hypotheses. This *loop-back* reflects the iterative nature of a PM analysis, where mining process can be refined, and new or different data can be integrated to get more efficient results [9].

In this work, the analysis is performed using pMineR [7], a software library in R specifically born to support PM analysis in healthcare and recently improved in the direction of exploring differences among the mined pathways.

2.1 *Dataset and Preprocessing*

We consider the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) dataset [5], that comprehends demographic and clinical longitudinal information of patients enrolled in 23 distinct ALS clinical trials. PRO-ACT includes a large sample size and high frequency of visits, that allow a precise characterisation of how ALS progresses in the study population.

To homogenize the data set, we selected the patients with a shared panel of exams and the variables with less than 50% of missing values. Then, we filtered out the subjects with unknown time of onset and the visits without a functional assessment or performed before the trial start. Finally, for each visit we converted the available functional evaluations into the Milano-Torino staging (MiToS) system [6], that is, 4 binary dynamic variables that trace when a specific functional domain among (1) Walking/Self-care, (2) Swallowing, (3) Communicating, or (4) Breathing is impaired. For each binary vari-

able, a value of 1 indicates that the domain is impaired, 0 is used otherwise, resulting in a string such as M_0001. We then converted the data into an EL, considering as events: the *disease onset*, the *trial start*, the *MiToS impairments*, and the *death* or the *censoring time*. For each event we coded as attributes the subjects' static information (sex, age at onset, site of onset) and the results of the lab tests (*e.g.,* hemoglobin, sodium, bilirubin).

This resulted in an EL of 9,009 logs and 31 attributes, that refers to 1,874 subjects.

2.2 *Direct Following Diagram*

The DFD is probably one of the most intuitive graphical languages which simply connects two nodes representing two events with an edge when they are subsequent in at least one trace. We generated a DFD for the whole EL. To prune the graph, different kind of thresholds can then be applied (*e.g.,* based on the absolute/relative number of transitions, timing, etc). In pMineR, the DFD is available by the class *firstOrderMarkovModel* and implements some additional features to allow the exploration of:

- *time to fly*: the kernel density distribution function of the time needed to move from a given node to a destination node;
- *survival functions*: Kaplan-Meier (KM) curves can be built, including possible constraints to select the cohort(s) (*e.g.,* passing or not through specific nodes and nodes playing the role of *censoring*), and then compared with a log-rank test to check statistically significant differences among the cohorts;
- *deltaGraphs*: two DFD graphs, each built for instance on a cohort with different baseline characteristics, can be overlaid to measure the differences in terms of transaction probabilities among nodes. Thresholds can be applied to reduce the noise and focus only on the relevant differences.

2.3 *CareFlowMiner*

From the DFD, the process is mined using the CFM algorithm, whose version implemented in pMineR is an extension of [8]. Starting from a *root* node, each trace in the EL contributes to create a branch of a tree where the top level (first after the root) represents the first event of each trace, and the next levels correspond to further events of each trace. Each node is labelled with the name of the corresponding event and additional information, such as the number of patients passing through that node or statistics about the time needed to reach it. To avoid the Spaghetti Effect, a CFM tree is normally pruned on the base of a threshold, to exclude highly infrequent paths and reduce the complexity of the tree. On the one hand, the tree tends to explode in terms of nodes and edges. On the other hand, the meaning of the language is easy to understand and the algorithm is clear. The latter, differently from existing PM algorithms such as Alpha Algorithm or Fuzzy Miner [9], helps in reducing the psychological barrier the clinician may have with respect to what they can feel as *black box* solutions. By itself, CFM is an algorithm able to show the most frequent paths, thus revealing strange behaviours and suggest further investigations, but it is not able to provide p-values or confidence intervals on their occurrence. pMineR overcomes this limit offering the opportunity to compare two CFM graphs (corresponding for instance to two different cohorts of patients, such as male vs female, young vs old, or with vs without a comorbidity at baseline). Such two CFMs are compared node by node with a Fisher's exact test or a chi-square test (depending on the cardinality of patients passing through the node) for dicotomic categorical variables, or with a Wilcoxon-Mann-Whitney test to compare the time needed to move from the *root* to the node or to move from the node to a given possible event (*e.g.,* death).

3 **Results**

Fig. 2 reports the exploratory analysis of the paths followed by the whole study population mined through the DFD. Notably, the topological organization of the DFD reflects the increasing trend of functional domains affected as ALS progresses: none at

the top (corresponding to the first trials' visit), one thereafter, and so on until the final states M_1111 (all domains affected), censored, or death are reached.

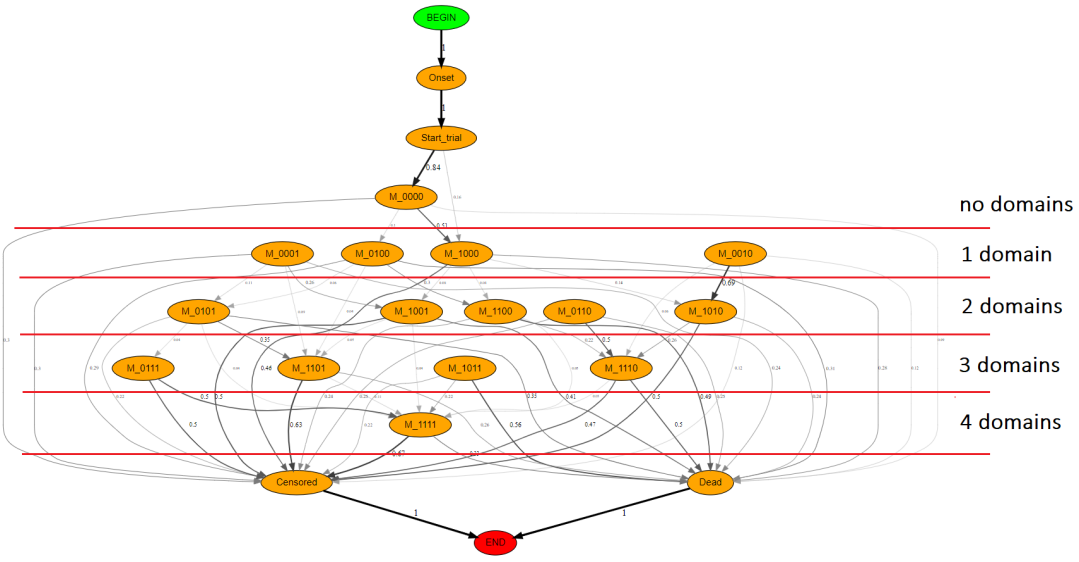

Figure 2: DFD graph representing the paths followed by the study population, delineating the increase in disability experienced by the subjects. Only the arcs with a transition probability > 0.03 are represented.

We then employed the DFD to shed some light into the kinetics of different clinical subtypes of ALS patients. Based on their site of onset, we compared the paths of spinal vs bulbar patients, under the hypothesis that their impairment patterns would differ. By stratifying on these cohorts and inspecting the corresponding DFD deltaGraphs reported in Fig. 3a), we can test the hypothesis by analyzing, for instance, the transition from M_0000 (first visit without an impairment) to M_1000 (impaired in Walking/Self-care), that results probabilistically different. The corresponding difference in terms of transition times can be inspected in terms of KM curves compared by a log-rank test (see Fig. 3b). It is of course pivotal to take into account the cardinality of the considered edges, to avoid results that overfit a very specific class of cases.

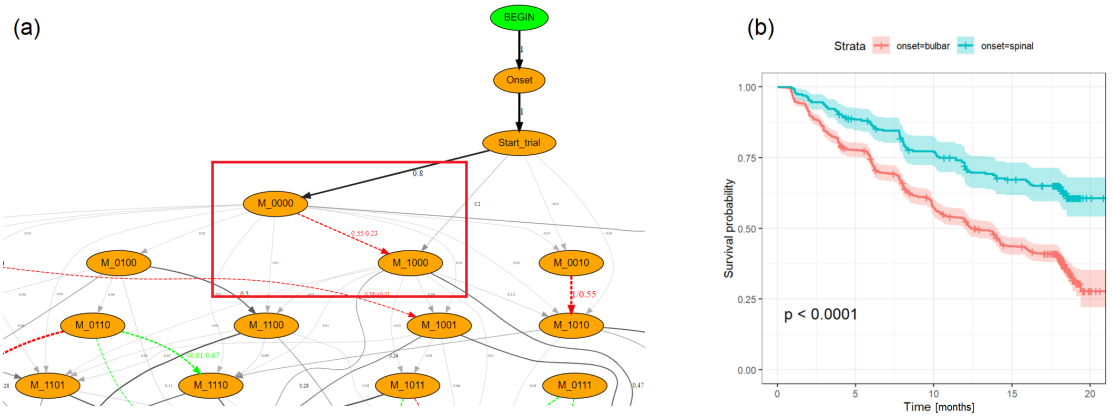

Figure 3: (a) Zoom on the DFD *deltaGraph* obtained stratifying the population by onset site (spinal vs bulbar). The highlighted edges represent an increased transition probability for the spinal (red) or bulbar (green) cohort, thresholded for displaying only differences between the probabilities greater than > 0.03. (b) KM curves of the time passing from M_0000 and M_1000 for the two cohorts. The log-rank test shows statistically significant difference between the cohorts.

Figure 4 reports the most frequent patterns mined through the CFM, here built starting from the node M_0000. For each node, the total number of patients passing through it (in brackets), as well as the minimum, median, and maximum time needed to reach that node from the root (second line, in days) are shown. The edges report the percentage of patients passing through the son node with respect to the entire population.

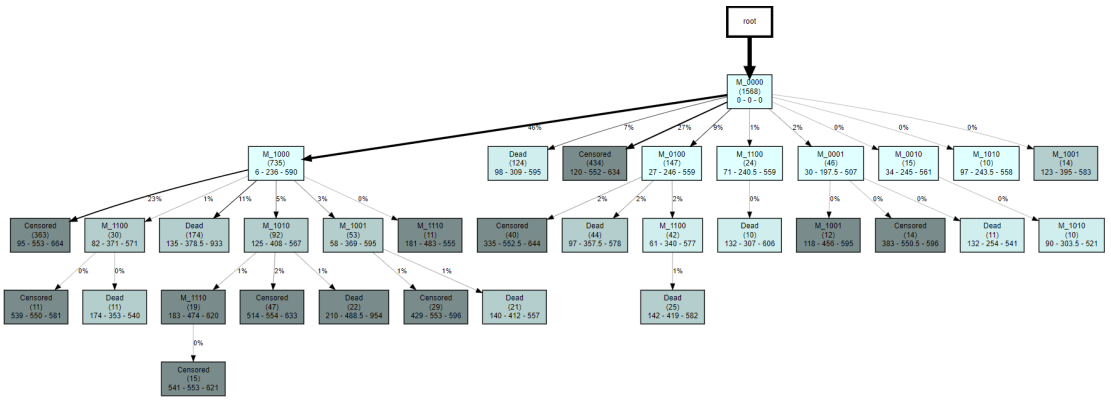

Figure 4: CFM graph built starting from M_0000. Each node reports the total number of patients passing through it (round brackets) and the min-median-max time, in days, needed to reach it from the root. Colours are graded on the median times, with intervals: <250, 251-350, 351-450, and >450 days.

As the DFD, the CFM tree can also be stratified by a variable of interest in order to assess any significant differences in the paths' occurrence according to the subjects characteristics. Guided by clinical hypotheses, we explored the distribution of the subjects in the nodes, focusing in detail on:

- age at onset (quantized into two levels according to its median value = 57 years) on the death, testing that older age at onset corresponds to a worse outcome,
- onset site (spinal vs bulbar) on the occurrence of the impairments, further checking that a spinal onset early affects motor skills while a bulbar one causes early dyspnea, dysphagia, or dysphonia.

Fig. 5 shows the obtained graphs. Each node reports the number of subjects passing through it for each cohort (young/aged onset or spinal/bulbar subjects, respectively), with the ratio in brackets, and the p-value of the Fisher's exact/chi-squared test, depending on the cardinalities involved in each node, on their distribution.

The results match with the expectation, quantitatively showing the exposure to early death for the oldest patients (the risk of death is significantly higher in most of the *Dead* nodes). It also emerges a significant predominance of a first impairment in the walking/self-care domain for the subjects with spinal onset (M_1000, ratio spinal/bulbar equal to 6.9), and a significant predominance of first impairment in the swallowing or in the communicating domains (M_0100 and M_0010, with a ratio spinal/bulbar equal to 0.14 and 0.36, respectively) for the bulbar onset subjects.

4 Conclusion

In this work, to support the investigation of the disease trajectories in ALS, we performed a PM analysis of a dataset of clinical trial patients. Mining the processes by two different algorithms, we outlined the impairments' patterns followed by the patients and inspected the effect of specific covariates on their probability and timing of occurrence.

The adopted approach allowed to formulate, to test and to iteratively refine hypotheses, also suggesting further directions of research, such as the analysis of the prognostic effect of additional covariates collected at baseline or in correspondence of specific events. The graphical representations of the mined process effectively supported the dialogue with the clinicians, helping in getting access to the information recorded in the data and in communicating the findings.

We believe this work can provide an original perspective in analyzing how ALS evolves. The mined processes can be exploited as decision support systems, indicating the probability of a patient to follow a given path based on his/her characteristics; further, they can allow to simulate the likely evolution of the disease and, in the future, to assess the impact of treatments. Future work will focus on validating the mined processes with real-world data, to test how general are the observed patterns.

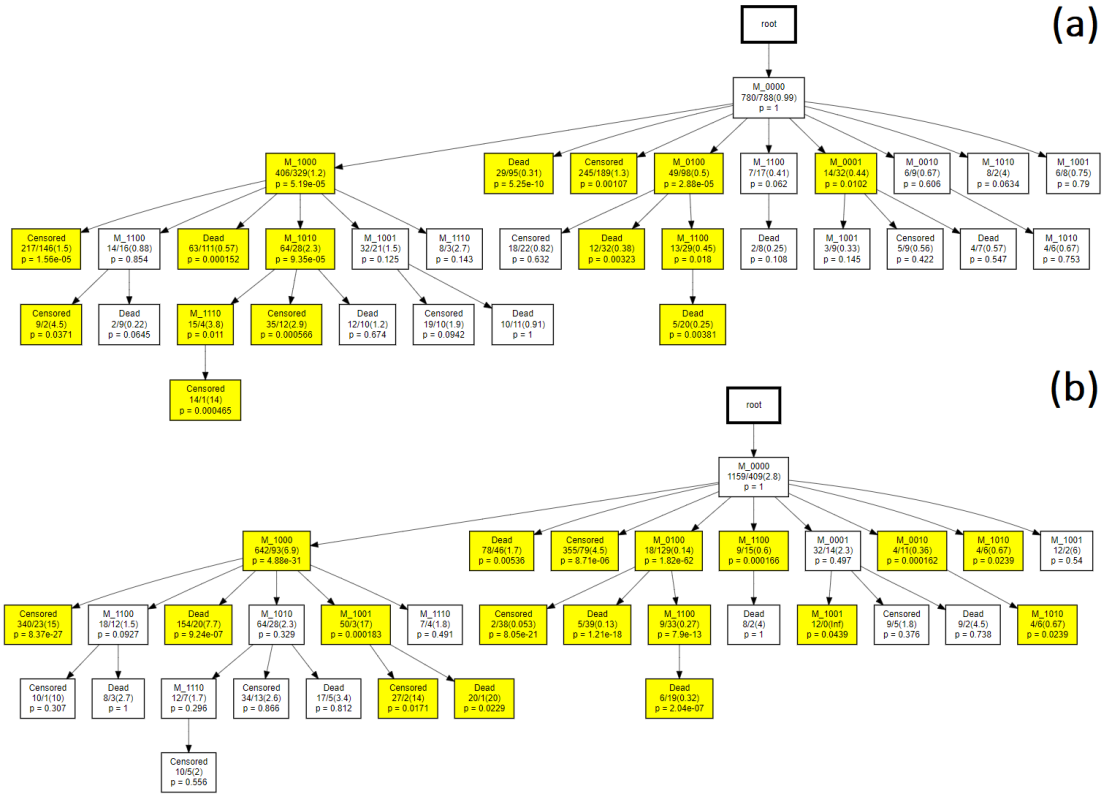

Figure 5: CFM graphs built starting from M_0000 and stratified for (a) quantized age at onset or (b) onset site. Each node reports the number of patients passing trough it for each cohort, the ratio of the two cardinalities (in the round brackets), and the p-value for the Fisher's exact/chi-squared test. The node box is coloured in yellow if the p-value is lower than a given threshold (here 0.05).

Funding

This work was partially supported by the University of Padova project C94I19001730001 and by the Italian Ministry of Health (Ricerca Finalizzata) grant RF-2016-02362405.

References

[1] E. Beghi, A. Chiò, P. Couratier, *et al.* "The epidemiology and treatment of ALS: focus on the heterogeneity of the disease and critical appraisal of therapeutic trials." *Amyotrophic Lateral Sclerosis* 12(1), 1-10, 2011.

[2] M. Müller, M. Gromicho, M. de Carvalho, S.C. Madeira. "Explainable models of disease progression in ALS: Learning from longitudinal clinical data with recurrent neural networks and deep model explanation." *Computer Methods and Programs in Biomedicine Update* 1, 100018, 2021.

[3] A. Carreiro, S. Pinto, M. de Carvalho, *et al.* "Classification of Clinical Data using Sequential Patterns: A case study in Amyotrophic Lateral Sclerosis." *2nd Workshop on Data Mining in Healthcare and Medicine, at SIAM International Conference on Data Mining*, 2013.

[4] A. Zandonà, R. Vasta, A. Chiò, B. Di Camillo. "A Dynamic Bayesian Network model for the simulation of Amyotrophic Lateral Sclerosis progression." *BMC bioinformatics* 20(4), 1-11, 2019.

[5] N. Atassi, J. Berry, A. Shui, *et al.* "The PRO-ACT database design, initial analyses, and predictive features." *Neurology* 83(19), 1719-1725, 2014.

[6] A. Chiò, E.R. Hammond, *et al.* "Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis." *Journal of Neurology, Neurosurgery & Psychiatry* 86(1), 38-44, 2015.

[7] R. Gatta, M. Vallati, J. Lenkowicz, E. Rojas, *et al.* "Generating and comparing knowledge graphs of medical processes using pMineR." *Proceedings of the Knowledge Capture Conference* 1-4, 2017.

[8] A. Dagliati, V. Tibollo, G. Cogni, *et al.* "Careflow mining techniques to explore type 2 diabetes evolution." *Journal of diabetes science and technology* 12(2), 251-259, 2018.

[9] W. van der Aalst "Process mining: discovery, conformance and enhancement of business processes." *Springer-Verlag* 2, 2011.