Air Pollution Profiling through Patient Stratification: Study of ALS Staging Systems Usefulness in Facilitating Data-driven Disease Subtyping and Discovery of Hazardous Ambient Air Pollutants

Notebook for the iDPP Lab on Intelligent Disease Progression Prediction at CLEF 2023

Mohamed Chiheb Karray[1]

[1]*Independent Researcher, Sfax, Tunisia*

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease characterized by heterogeneity in disease progression and short lifespan after symptoms onset. The longstanding heterogeneity problem is still driving the research community to identify robust ALS disease subtypes through different approaches. Unsupervised machine learning techniques are among the sought-after computational tools to help discover reliable subtypes. The reproducibility of studies employing automated computational tools heralds a promising future for precision medicine where discovery of groups of similar patients is of paramount importance for clinical trials, personalized treatment plans as well as early disease diagnosis. However, the current state of the art of using machine learning techniques for ALS has started to make many researchers from the community address the problem of low-quality, hard to validate conducted studies. At the same time, accumulating set of findings is associating short and long-term exposure to air pollution with increased risk of having ALS or ALS disease worsening. In our study, we turn to an overlooked ALS prognosis tool to stratify patients using clustering techniques. A subsequent data profiling process made us identify interesting air pollution profiles that corroborates findings about some air pollutants being a risk factor for disease onset or disease worsening. We have also studied the impact of ingesting machine learning-based classification and survival analysis models with environmental data on the performance of such models. The targeted task through which such models were assessed is the prediction of occurrence of ALS-related events characterizing an endpoint of the disease or the reaching of a serious disease stage.

Keywords

Amyotrophic Lateral Sclerosis, Patient Stratification, Clustering, ALS Staging Systems, Air Pollution, Risk Prediction

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease with an incidence rate ranging from 2 to 4 per 100000 people. The incidence rate is subject to increase in coming years [1]. The life span after disease onset is typically between 3 and 5 years but could be lower for patients with faster disease progression. Adding to that, the diagnosis delay can go from 12

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

chiheb.karray@ieee.org (M. C. Karray)

months to 18 months in general [2] but could be even longer reaching 27 months in some cases [3, 4]. The incurable disease is characterized by a high degree of heterogeneity making it hard to diagnose[5]. The clinical heterogeneity of ALS is noticeable in terms of age at disease onset, site of onset, disease progression rates as well as the manifestation of behavioural and cognitive changes [2]. Duration of diagnostic delay is another heterogeneity factor of the disease [6]. The variability of the disease patterns is a long-debated problem in ALS research which is hindering the quest for cures that could stop or at least slow the progression of the disease for more than few months. This multi-factorial variability is considered to be the "greatest challenge" for ALS researchers [5]. Tackling the problem of finding reliable ALS subtypes by means of clustering tools is a way to help researchers design more personalized treatment plans and more precise diagnosis and prognosis tools.

Throughout our participation at iDPP 2023 competition and knowing that ALS patient stratification is among the goals of the "Brainteaser" project, we have attempted to design a considerably reliable clustering process whose outcome is provided in the greatest possible details for clinicians to analyze and make suggestions for future improvements in terms of expected characteristics of clusters. Being aware of the importance and the recency of studying relationship between air pollutants and ALS, we thought that clustering patients would certainly help us study the potential added value of environmental data on disease progression-related tasks such as the ones targeted by the iDPP 2023 competition [7, 8].

That being said, our paper is organized as follows: Section 2 introduces related works; Section 3 describes our approach; Section 4 discusses our main findings; finally, Section 5 draws some conclusions and outlooks for future work.

2. Related Work

2.1. ALS and exposure to air pollutants

Exposure to particles carried by air has been pinpointed to be a risk factor for neurodegenerative diseases and ALS is no exception. Ambient air pollution is found to be among the plausible causes of oxidative stress and neuroinflammation which could lead to neurodegeneration[9, 10]. According to many studies, ambient particulate matter (PM) is among the main culprits behind neurotoxicity [10]. The smaller the particulate the more fatal they become [9]. We can distinguish between PM_{10} or coarse particles (with diameter $<$ 10µm), $PM_{2.5}$ or fine particle (with diameter $<$ 2.5 µm), PM_1 submicron particles (with diameter $<$ 1µm) and UFPM/$PM_{0.1}$ ultrafine particles (with diameter < 100 nm) [9, 10]. Particulate Matter in general is a combination of chemical and biological particles that differs by regions, seasons, weather patterns and local pollution sources [9, 11]. In ALS literature, the few studies [12, 13, 14, 11, 15] that approached the effect of air pollutants on the disease aggravation have focused on studying coarse PM and fine PMs for the PM category among other air pollutants such as Ozone O_3 and Nitrogen Dioxide NO_2. The aforementioned studies have mainly studied long-term exposure impact and to a lesser extent the impact of short-term exposure. All the covered studies have conducted association studies between levels of exposure to one or more air pollutants and the aggravation of ALS. This aggravation was sometimes measured through "surrogate" indicators like the

admission to hospital for ALS being a primary admission reason [13, 11]. Conclusions of studies investigating the relationship between pollutants and ALS aggravation (or Risk) were contrasted on whether a given particle is associated with ALS risk or not. Nitrogen Dioxide from all the studied air pollutants seems to be a risk factor of ALS on which many studies have agreed [16, 13]. For PM_{10} and O_3, consensus was far from being reached on them being behind increased risk of ALS. As for $PM_{2.5}$, covered studies are mostly agreeing on their hazardous impact pertaining to ALS aggravation/risk of occurrence. The subtilty about $PM_{2.5}$ is that its composition and not its concentration level within a region is what makes it harmful and thus coinciding, at higher odds, with people in a region being impacted by ALS [16, 11].

2.2. Clustering for ALS patient stratification

Unsupervised Machine Learning techniques such as clustering are helpful in finding groups of similar patients according to their characteristics (demographic, clinical, labs data, genetic, imaging-derived ...). The advantage offered by such techniques is that they are data-driven. Such techniques spare the clinicians the burden of 'manually' dissecting into separate groups the patients, usually characterized by an important number of features, through the use of criteria derived from the literature or inspired from their own medical experience. For diseases such as ALS, characterized by high heterogeneity, the use of clustering techniques could be much of a help to uncover clusters or 'profiles' of patients among a studied population which might be indiscernible by a human expert (clinician). However, the improper use of data-driven techniques and the adoption of their outcomes, if applied on commonly used datasets in a given study field, could have catastrophic consequences. In the ALS community, adding to the issue of lack of use of such techniques, outcomes of the few studies relying on clustering for patient stratification have never been quantitatively examined.

During the last few years, some researchers have been voicing their discontent about the improper use of clustering techniques for ALS patient stratification [1, 17]. For example, the very recent review on ML-related stratification and prediction models for ALS [17] has concluded that data-driven stratification techniques for ALS lack validation as well as reproducibility. In [1], they have contested the design of clustering studies as a whole.

In fact, to ensure that clustering results are reliable, many criteria should be satisfied. The most important of them is the choice of a similarity metric that reflects domain expertise in relation to what makes patients close to each other in terms of data characteristics[18]. The second consequential step is to ensure the absence of distance concentration phenomenon with respect to the chosen metric [19, 20]. That is to assess whether data representation of patients paired with a chosen similarity metric would inherently lead to distant and separable non-interleaving groups. Testing for the absence of a distance concentration phenomenon for the chosen metric is a statistical guarantee of the data clusterability [21] especially for high-dimensional data. The not less important step of choosing the clustering algorithm itself is detrimental to the outcomes to be obtained since the whole progression/diagnosis process will be built on such outcomes. In a context where data tend to form clusters of different shapes and densities with an absence of a unanimous agreement on possible number of clusters to obtain; utmost care should be given to the choice of a clustering algorithm. Finally, the internal validation of the clustering outcome which is generally an index that assesses the

quality of obtained clusters has to also encode the reality of the domain expertise. Internal clustering validation techniques generally tend to say whether a chosen/produced number of clusters is correct compared to other partitions (stratification) with different numbers of clusters. Such 'correctness' differs from one validation method to another and depends on the way the validation method defines 'dense' and 'separate' clusters. Another overlooked problem in clustering validation is the inadequacy of chosen validation metrics to the applied clustering algorithm. Even in the ML community of clustering validation, neglecting one or more of the above-mentioned steps is rampant. The problem itself is not drawing so much attention within the ML community since the inception of the second spring of Neural Networks (a.k.a Deep Learning and Differential Programming) a decade ago.

Even though problems pertaining to clustering validation are not fully solved or adequately addressed, the last thorough review of internal clustering validation techniques dates back to 2013 [22]. As for external stratification validation, a better performance of a disease progression model could be thought of as a surrogate for a better stratification of patients. However, in a brief preliminary study, we have shown that a better performance of a disease progression model predicting ALSFRS-R slope does not entail a better clustering quality [23]. For the same study we have also managed to slightly improve the disease progression model by applying the same Random Forests model on clusters of patients validated by an Ensemble of Internal Clustering validation metrics. For the 2015 ALS patient stratification challenge, for the same task of disease progression, the winning model used patients as a single group outperforming many submissions having used a varied number of cluster numbers and different clustering approaches [24]. Quality of obtained clusters for different participating groups were never assessed to the best of our knowledge.

In[25], authors have used K-Means as a clustering algorithm and the Euclidean distance as similarity metric (default choice even if it is not mentioned). For ALS patient stratification, given the approach followed by the study, different problems will arise. First, the Euclidean distance is not adapted to mixed-type data (where categorical and continuous data are present). Then, K-Means (an instantiation of Gaussian Mixture Models), assumes that clusters are normally distributed around a mean so that clusters tend to be organized in concentric circled (in 2D) with similar number of patients existing in each cluster (K-Means uniform effect[26]). Afterwards, simply executing the clustering algorithm for different cluster numbers (ranging from 2 to 15) and concluding that a given cluster number is the most adequate one according to the K-means cost function value for different cluster numbers is fallacious. K-Means as an algorithm is not capable of returning the whole dataset as a single cluster if data are not clusterable. Thus, any given cluster number will provide the user with a clustering. Afterwards, the used t-SNE projection in the study to demonstrate the quality of the stratification clearly contradicts the claim of the adequacy of the clustering method to the used data. In fact, the provided projection proved that the used clustering algorithm has divided the dataset to balanced sets of patients even for the case of a cluster that should be taken as is without further divisions. Similar examples are widespread in the data-driven ALS patient stratification literature.

Using a neural network (a feed-forward neural network or an autoencoder) with a differentiable K-Means loss function would have the advantage of learning a new data representation (with a lower dimensionality) but will not palliate with the inherent assumption of the K-Means

loss function. That is to say that patient stratification using clustering techniques within the ALS research community is plagued with a lot of wrong design choices.

3. Methodology

3.1. Feature Engineering

3.1.1. Features obtained from static Data

The iDPP dataset is characterized by a high number of variables with missingness rates making them unexploitable for data analysis (percentage of missing value >65%). In our study, we have deemed that some features even if found highly missing among the studied population could be informative if grouped with features with the same semantic. Henceforth, eight features related to different diseases recorded were grouped under a single categorical feature dubbed 'Disease_History'. For the created feature, we have found more than 75 different combinations of diseases. Another related feature counting the number of diseases for each patient was also created ('Number_of_Diseases'). The second feature was thought to be helpful in profiling groups of patients disregarding the details related to their disease history. In the same line of thought, features encoding the presence of ALS-characteristic gene mutations were transformed since they were fragmented over two centers (Turin and Lisbon). We also thought of the possibility of finding regional clusters of ALS especially when augmenting the static and visits data with environmental data. Therefore, a feature called 'Nationality' was created using the presence of values in variables with suffixes Turin and Lisbon.

Having in mind that our objective is to better profile different groups of patients showing similar disease progress patterns, we divided information carried by the 'onset_limb_type' into separate groups of anatomical locations. The groups were Upper/Lower, Right/Left and Proximal/Distal. Proceeding with the encoding of the anatomical groups made us discover a set of instances where the anatomical description of the affected limb was present while 'onset_limb' was set to False. The diagnosis delay being the difference between diagnosis day and disease onset date was also included in our feature set extracted from static data. As for features related to height and weight, we have used them to compute two features 'BMI_Before_Onset' and 'BMI_at_Onset'. The rest of features whose missingness rate exceeds 65% were deleted. And to make the set of static data features exploitable by ML algorithms we have proceeded with a simple replacement of missing values with the same constant value.

3.1.2. Features obtained from visits data

ALS staging system outcomes are believed to serve as a "potential input" to patient stratification methods according to [1]. Consequently, in order to uncover the heterogeneity of patients belonging to the studied population we considered relying on different staging systems for ALS. MiToS [27], Fine till 9 [28] and D50 [29, 30] staging systems were employed. The King's staging system requires the presence of an information about the introduction of gastrostomy for ALS patients. Usually, this information is known if answers to item 5 of the ALSFRS-R questionnaire were present with two suffixes (A and B) [31]. For our case, this type of information, namely

PEG (Percutaneous Endoscopic Gastrostomy), was used as an event to be predicted. Absence of the aforesaid information made the use of King's staging system impossible for our case.

For the three employed staging systems, a common set of pieces of information could be obtained which is the set of stages reached by each patient. MiToS and Fine till 9 could provide us with a more dynamic description of the disease progress for every new visit. Whereas the D50 staging system being based on a parametrized exponential disease progression model would provide a possible staging that could be optimistic, pessimistic or even misleading (in case of lack of visits data where their corresponding ALSFRS-R are not belonging to two different ranges at least). However, with enough data adhering to the requirement of the model we can get a standardized view of disease progress of patients even with different disease progression dynamics. The second set of details that could differentiate patients is the set of affected domains or functional groups being attained at each visit. This type of information is only provided by MiToS and Fine till 9 staging systems. The other reason for using two staging systems other than D50, is the difference between the aforementioned systems regarding their "sensitivity" to different phases of ALS progression. At an early phase of the disease progression, MiToS tends to be less sensitive to changes occurring to the patient while Fine till 9 was reported to be more sensitive to changes taking place at the same early disease progression phase [28, 32]. Differences between both systems could help us identify fine-grained groups of patients with heterogeneous progression patterns.

Features extracted from visits data were categorical (nominal and ordinal) and continuous. The information carried by engineered features falls under one of the following categories: Affected regions, reached disease stage, duration of each stage, disease progression severity and frequency of visits. For MiToS and Fine till 9 we distinguish two types of affected zones for a patient. MiToS looks for affected functional domains "that involve loss of autonomy" [33] while Fine till 9 helps at identifying affected functional groups (bulbar, fine motor, gross motor and respiratory). For each patient and after running MiToS and Fine till 9 staging algorithms, we kept affected zones for his initial and last visits. Each sequence of affected zones constitutes a unique category so that a pair of initial and final set of affected zones would be able to differentiate patients from each other. Staging systems also enable the quantification of the reached disease stage for each newly fed ALS visit data (answers to the ALSFRS-R questionnaire and the total ALSFRS-R score). With the availability of a such information, the sequence of reached stages for each patient could be generated. Obtained sequences of stages could identify disease reversals and cases where the disease progression is stable/ized. Obtained sequences were simplified so that a patient with a stable stage for successive visits would have a single stage taken into account for such visits. Coupling the inferred sequence with visits times, the duration of each stage for each patient is computed. Once durations are computed, we recurred to normalizing them with respect to the overall disease duration. The caveat in determining the normalized disease stage duration is the absence of a disease endpoint in training data (death or tracheostomy)[34]. To circumvent the absence of a unique disease endpoint for patients (in the training set), we assumed that the endpoint is the prediction horizon chosen for the competition. Consequently, the stage reached by a patient during his last recorded visit is supposed to last until the prediction horizon. Our assumption followed the same adopted approach by the competition organizers to compute the ALSFRS-R slope. The slope was computed by supposing

that a 'virtual visit' took place at the date of disease onset and thus the perfect ALSFRS-R score remained unchanged until the occurrence of the first 'real' ALS visit. The principle of assuming the absence of any changes in disease state with the absence of real data remains questionable especially for a disease characterized by high heterogeneity and short lifespan after onset. Such an issue is to be briefly discussed in the results section.

For the third staging system used in our study, D50, we have kept two types of information: the date at which the patient is expected to lose 50% of his functional abilities (or D50) and the final standardized stage reached by the patient according to rD50 values (visits dates transformed using the D50 inferred parameters for a given patient). The final standardized stage is determined using experimental intervals from the literature. According to [35], a patient could be in one of 3 stages: early stable phase of progression ($0<$rD50<0.25), early progressive phase of progression ($0.25<=$rD50<0.50) and late progressive and stable progression phase ($0.5<=$rD50).

To represent the disease progression speed, we have computed the early and late ALSFRS-R slopes, where the early slope is rate of ALSFRS-R points lost by month between the disease onset and the date of the first visit and the late slope is the same rate compute between the last visit and the first one. Both quantities were negative (in contrast to the slope variable present in static data of patients). Categorizing the speed based on the computed slope was intended as a next step to transform the slopes, however the literature was not conclusive on clear-cut thresholds to map slopes to progression categories given certain time periods. The other issue with slopes in our case is the high variability of time before the first ALS visit.

3.1.3. Features extracted from environmental time-series

Static representation of different times series of environmental particles was preceded by two main exploratory data analysis procedures: count of environmental data presence for each particle for the patients and visualization of time-series along with ALSFRS-R curves. We have noticed that not all patients have environmental data. For the particles whose data are the most present among patients, only 26% of patients belonging to datasets A, B and C do have such particles-related data. Those particles are the atmospheric/weather-related particles. The presence of pollutants data is highly variable ranging from 3% to 21%. Looking for patients having data related to all particles at once was fruitless: none of the patients belonging to the training set had records of time-series pertaining to all 15 environmental particles covered by the iDPP dataset. The objective of the competition is to investigate the impact of the exposure to pollutants on improving the performance of models predicting risk of medical events. As a result, we focused our preliminary analyses on studying the presence of pollutants data among patients as well their patterns with respect to disease progression. In addition to studying the presence of each pollutant data independently of other pollutants, exploring the number of patients having records of two pollutants out of seven led to the necessity to remove the least occurring ones. For example, only two patients in the training set (all datasets included) had data on C_6H_6 and NO_2. The most frequent joint occurrences are of (PM_{10},O_3) with 234 patients and ($PM_{2.5}$,O_3) with 188 patients out of 467 patients.

Visualizing time-series of the most frequent pollutants such as PM_{10}, $PM_{2.5}$ and O_3 (for patients having records of one of the three pollutants) along with the ALSFRS-R curve led

to noticing patterns such as: the absence of time series data for a large period starting from the date of the first ALS visit, the ALSFRS-R records do exist before/after the presence of any information on the pollutant (both curves do not share any common time periods) … That is to say that adding to the irregularity of ALSFRS-R entries, pollutants time-series data are not always useful to study correlation between high exposure and disease progress let alone to study causality between exposure to one or more pollutants and disease worsening. The previous visualization procedure was mainly performed on patients with fast disease progress patterns.

In order to avoid a high rate of missing data for the features representing environmental data, we had to go for the most frequently present pollutants, namely PM_{10}, O_3 and $PM_{2.5}$. The European Air Quality Index chart (EAQI), provided us with different levels of exposure to each of the three particles. Those levels were used to devise two features for each pollutant: *ratio_medium* and *ratio_poor*. *Ratio_medium* represents the ratio of days during which a patient was exposed to medium concentrations of a pollutant over the total number of days of exposure. *Ratio_poor* stands for the sum of ratios of exposure to high, very high and extremely high concentrations of a pollutant. Adding to the couple of features mentioned beforehand, we added the number of exposure days, and basic aggregate statistics about pollutant time-series such as Min, Max and Mean concentrations of pollutants based on the 'measurement' feature present in the competition dataset. Presence of different combinations of the most frequent pollutants among the set of patients was also included in the form of Boolean features quantifying the joint existence of environmental data (presence of (PM_{10}, $PM_{2.5}$, O_3), presence of (PM_{10}, $PM_{2.5}$), presence of (PM_{10}, NO_2, O_3) …).

3.2. Identification of patient clusters and environmental profiles

3.2.1. Patient Stratification

To be able to extract clusters of patients we have only relied on features extracted from visits data with the exception of the diagnosis delay obtained from static data. Since the data to be clustered was a mix of categorical and continuous attributes we chose Gower distance as a similarity matrix. We have checked for the absence of the distance concentration phenomenon using the distribution of distances. Then we applied t-SNE for dimensionality reduction and data visualization. Patients' data represented by the obtained t-SNE embedding was clustered using the HDBSCAN algorithm[36]. The algorithm was used for its capacity to extract clusters with varying densities (and shapes) without the need to provide it with a pre-specified number of clusters to be obtained. The stratification process was followed by a detailed profiling of each obtained cluster (Tables A.1, A.2 and A.3).

3.2.2. Pollution profiles of stratified patients

It can be clearly understood that features extracted from static data and environmental time-series were not used for the clustering of patients. Having quantified exposure rates according to pollution levels specified by the EAQI, we have used them to create pollution profiles. A pollution profile is characterized by a collection of features that provide an overview of air pollution exposure for a given set of patients. For our case, that profile is given by poor and high

rates of exposure to 3 different pollutants: PM_{10}, $PM_{2.5}$ and O_3. This exposure is measured by the ratio of number of days of exposure to medium and higher concentrations of a particle over all the exposure duration to the same particle. Information of such ratios are carried by features <Particle_Name>_ratio_Poor and <Particle_Name>_ratio_Medium.

To measure the similarity between clusters in terms of exposure to a certain pollution level of a pollutant, we have estimated the Probability Density Functions of pollution ratios of each particle within each cluster of patients. Then, we have employed the Jensen-Shannon divergence to obtain a similarity matrix for each particle and each level of exposure for all the clusters. The next step was to group Patient Clusters given the computed similarity matrices. Given that the overall number of clustering operations is equal to the number of particles times the number of exposure levels (Medium and High), a co-occurrence matrix was computed. It will represent which Patients' clusters had similar exposure 'footprint' with respect to different pollutants and exposure levels. That co-occurrence matrix will serve as a new output to another clustering process to obtain the sought pollution profiles. For all the clustering operations, HDBSCAN was used as a clustering algorithm. The previously described process is inspired from the way of performing ensemble clustering.

3.3. Risk Prediction Models

Risk prediction for the 3 provided datasets was cast as a survival analysis task and as classification task. For the classification task, we have relied on an Ensemble of Classifiers where their number, their weights, their algorithms and their parameters were determined by an Automated Machine Learning method [37]. The ensemble was validated through a 5-fold cross validation process maximizing the accuracy score of the ensemble classifier. Two separate Ensemble classifiers were retained. Each one of them was trained on a different data representation (data with Environmental Features and data without Environmental Features). The time of the automatic search process was scheduled to be the same for both ensembles. As for the survival analysis model, we have relied on Survival RandomForests. The hyperparameter search process looked for the optimal set through the maximization of the C-index of the trained model. The same model was applied for data with and without environmental features.

For both scenarios, the hyperparameters search process was executed on a single dataset: dataset A for the Survival Analysis case and dataset B for the Classification case. Models with the retained set of hyperparameters were employed to predict outcomes of the rest of the datasets. For the classification case, the Ensemble was trained once on dataset B and predictions for tasks A and C were based on 'knowledge' acquired from data of task B. That is to say, that for the classification scenario we have performed a basic transfer of learnt 'knowledge' from task B that was used to make predictions for the two remaining tasks.

We have also made some initial attempts to train our models on obtained clusters of patients but the preliminary results were not encouraging due in part to the small size of some clusters (especially for the case of patients with environmental data). We have also paid attention to the important class imbalance phenomenon for the different tasks. For the class imbalance issue, we have tried a two-process data augmentation scenario. Its first step consisted in under-sampling the majority class. The next step was to generate synthetic data for both classes using

deep-learning based methods (2 variants of Variational Autoencoders [38]). The obtained data (original and synthetic) were merged and obtained models were applied on the blended 'new' dataset. Again, the preliminary results led to a decrease of the performance of survival and classification models compared to the sole use of the original data in the training process. Little to no feature selection procedures were performed during our experiments.

4. Results

4.1. Patient Stratification Results

The proposed clustering procedure led to obtaining 15 clusters and a cluster of outliers. From the t-SNE projection of the data where clusters are distinguishable by different colors (Figure 1), we can see that there are 4 main groups of clusters. Further details about the general characteristics of each cluster are provided in Tables A.1, A.2 and A.3 of appendix A.

Figure 1: t-SNE visualization of clustered patients

Clusters to left side of Figure 1 are clusters where we have patients with a unique visit but with different characteristics. Clusters at the center of the figure (Clusters 11, 12, 13 and 14) are clusters with the most noticeably rapid disease progression patterns.

The blue cluster (labeled as "Cluster 1") at the uppermost right side of the figure is a cluster of patients with very slow to constant disease progress rate (in terms of early and late ALSFRS-R slopes). In fact, it contains the highest number of people with constant disease progress after the first ALS visit (for the whole study population and for the population of patients with environmental data). Out of 174 patients (belonging to the whole study population) with late ALSFRS-R slopes equaling zero, 62 patients belong to "Cluster 1" which is equivalent to 35% of the patients with constant disease progress after the first ALS visit. The cluster with the second highest number of patients with constant progress is cluster "Cluster 3" situated below "Cluster 1" in figure 1. It contains 16% of patients with constant disease progress. Groups labeled as "Cluster 7" and "Cluster 6" come next in line with 15.5% and 12.6% of patients with constant progress respectively. Among the most prominent differences separating "Cluster 1" from clusters 3, 6 and 7 is the last Fine till 9 reached stage. For the first cluster 100% of patients

have stayed at stage 0 of the aforementioned staging system. Without any exception patients in clusters 3, 6 and 7 have reached stage 1 of the Fine till 9.

Coming to clusters visible at the exact center of the figure, "Cluster 11" is the most interesting one in terms of speed of disease progression with ALSFRS-R slope passing from -0.81 to -1.89 and a D50 stage of 2 for all the patients. Worse than patients of "Cluster 11", are those of "Cluster 13" with a mean ALSFRS-R slope decreasing from -0.75 to -2.2. Interestingly enough fast progressors of "Cluster 13" have a very similar pollution profile with the slow progressors of "Cluster 6" according to our results. Details about both pollution profiles will be provided in the next section.

"Cluster 12" is characterized by interesting properties in terms of progression speed as its mean ALSFRS-R slope goes from -1.4 to -2.58. More than half of the patients of this cluster have reached an advanced Fine till 9 stage (stage 4) and the rest of them reached stage 3. In 32.9% of the cases, patients belonging to this cluster have lost 3 functional groups and 8.2% of them have had their 1st ALS visit with 4 lost functional groups. This clearly entails that those patients have reached stages 2 and 1 long before their first visit. Consequently, if we want to extract time of each stage for such patients we wrongly obtain a mean standardized duration of 65% at stage 0 and near 5% for stages 1 and 2. Such obviously misleading durations are due to the fact of considering that patients had the perfect ALSFRS-R score since disease onset until their first ALSFRS-S visit. Even if such absence of data would probably not occur in clinical trials, coming with methods to estimate approximate dates of transition from stage 0 to stages 1 and 2 in such cases would make both stratification and disease progression models more reliable. D50 progression model will not be capable of filling the gap for various reasons. Among them we can cite the most important problem of subjectivity of the ALSFRS-R scale, the constraining need to provide the model with data points from different ALSFRS-R ranges and his tendency of nonlinear decrease which will not take into account cases of reversals (if they are really happening), improvement of functionality of different regions being covered by the ALSFRS-R scale …

Coming now to the cases of ALS reversals, we have detected 47 over the 1787 patients (2.6%) belonging to the pool of patients in training data for the different tasks subject of the competition. Over 21% of the reversal cases belonged to "Cluster 6", 19% of them were part of "Cluster 2" and over 12% belonged to "Cluster 1". With respect to MiToS and Fine Till 9 staging systems, they have indicated cases of reversals for 6 and 16 cases respectively. For one case, Fine till 9 returned a reversal from stage 3 to stage 0 (patient "0x73a002c9518c344906b2929c4a299cf9"). For the mentioned patient, his ALSFRS-R score has increased from 35 to 46 according to his visits data in less than 4 months. As for MiToS we have found only one noticeable case for a patient who went back from stage 2 to stage 1 (patient "0xabfbe3b27f02e837e604375dfddb6978"). For the same patient, his last Fine till 9 stage was 4.

4.2. Environmental profiles of stratified patients

Applying the method discussed in the methodology section led us to obtaining 3 clusters of pollution profiles and 1 cluster of outliers. Patients' "Cluster 10" was among the outliers. In fact, the computed Jensen-Shannon dissimilarity matrices have shown us that the pollution profile

of "Cluster 10" was the most distant from the rest of profiles. It is characterized by the lowest exposure rates to PM_{10} and $PM_{2.5}$. Profiles of Clusters 4 and 8 were different from the rest of the profiles but were less distant than "Cluster 10". Cluster 4 pollution profile is quite specific in terms of the mean rate of exposure to high concentrations of $PM_{2.5}$. As a matter of fact, it is the only cluster where patients were exposed to high concentrations of $PM_{2.5}$ particles for the longest period of time (almost 30% of 245 days in average). Patients of "Cluster 12 "had been exposed to high $PM_{2.5}$ concentrations in a very similar way as "Cluster 4". Even though pollution profile of "Cluster 12" had more similarities with 'clusterable' pollution profiles more than the rest of outliers, it was also labelled as an outlier. Yet, it was almost equal to 'high profile' Clusters 11 and 13 for the rate of exposure to medium concentrations of $PM_{2.5}$.

None of the pollution profiles of the 15 clusters had a perfect similarity for the probed pollutants PM10; $PM_{2.5}$ and O_3. However, we have had interesting groupings of pollution profiles of Clusters 1 and 14 from one side and Clusters 6 and 13 from the other side. Clusters 1 and 14 shared similar pollutant exposure profiles for medium concentrations of PM_{10} (Medium and High concentration levels) as well as $PM_{2.5}$ High concentration levels. Both clusters had the same distribution of exposure levels to medium PM_{10} concentrations. As for Clusters 6 and 13, they were characterized by similar exposure rates to Medium and High PM_{10} concentrations as well as similar exposure to Medium O_3 concentrations. They were dissimilar regarding their exposure to high and medium $PM_{2.5}$ concentrations.

4.3. Disease risk prediction results

As we have explained in the methods section, the outcome of the patients clustering was not exploited for the models we have used to generated our submitted results. In the present section, we will try to describe and comment different aspects of the performance of our models for the different tasks.

Figure 2, depicts obtained C-index scores for the different models submitted to the competition. In x-axis, 'CE' stands for Classifier Ensemble and 'SuRF' stands for Survival Random Forests.Each model's type is followed by the task name. Blue bars stand for models using static and visits data only and the orange ones represent models which are using the environmental in addition.

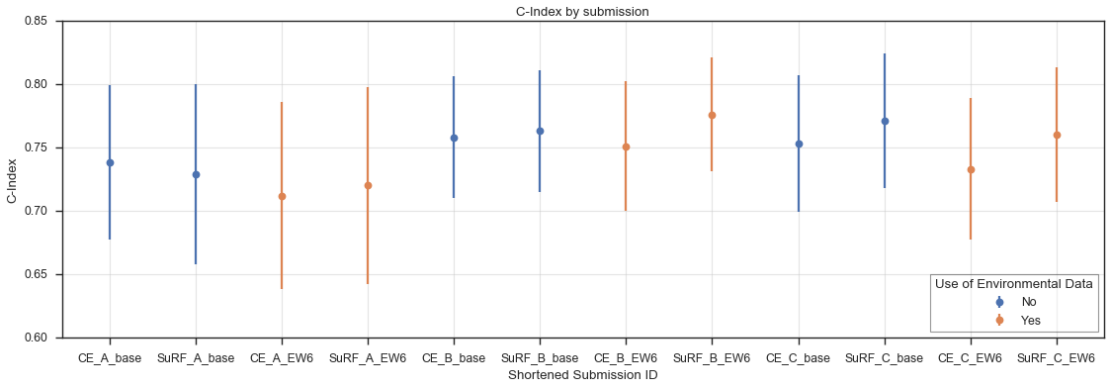

Figure 2: C-index with 95% confidence interval

From a survival analysis point of view, in (Figure 2) we notice that models trained using the ensemble classifier have managed to narrowly compete with Survival Analysis models especially for tasks A and B where we do have two competing events to predict. For task C, Survival Random Forests have better performed that the ensemble of classifiers. Knowledge gained from task B through the Ensemble of classifiers managed to make very similar performances in tasks A and C especially when uniquely relying on staging and static features.

Adding the environmental information to the survival model made it perform better than the ensemble classifier variant trained on three data modalities (static, visits and environmental). For task B, the gap between the survival model trained on all data modalities and the classifier trained on the same set of data is considerable. In fact, the model 'SuRF_B_EW6' is the best scoring among all models, all tasks included. The most important outcome to be drawn from Figure 2 is the positive influence of adding the environmental features on the survival model performance. To the contrary, the environmental data modality has negatively affected the performance of the ensemble classifier. While we have used two different configurations and set of hyperparameters to train the ensemble classifiers on the two different data representations (with and without environmental data) the outcome has yet to be meticulously analyzed. In the best-case scenario, while having been trained on the dataset of task B ensembles performed nearly the same for the two data representations. Table 1 offers a quantitative overview of different models performances in terms of C-Index.

Table 1
C-index of submitted models for subtasks A, B and C

Subtask	Use of Environmental Data	Submission ID	C Index & 95% Confidence Interval
3-A	No	CE_A_base	**0.738 [0.677, 0.799]**
		SuRF_A_base	0.729 [0.659, 0.8]
	Yes	CE_A_EW6	0.712 [0.637, 0.786]
		SuRF_A_EW6	0.72 [0.643,0.798]
3-B	No	CE_B_base	0.758 [0.71, 0.806]
		SuRF_B_base	0.763 [0.714, 0.811]
	Yes	CE_B_EW6	0.751 [0.7, 0.802]
		SuRF_B_EW6	**0.776 [0.73, 0.821]**
3-C	No	CE_C_base	0.753 [0.698, 0.807]
		SuRF_C_base	**0.771 [0.717, 0.824]**
	Yes	CE_C_EW6	0.733 [0.677, 0.789]
		SuRF_C_EW6	0.76 [0.708, 0.813]

Table 1 shows that the prediction models did not manage to get their performances improved by adding environmental data except for subtask 3-B. For the aforementioned task, the survival model had a slight performance increase from 0.763 to 0.776 in terms of C-index when environmental data was added.

We will now analyze performance of submitted models from a classification perspective. For task A, as we can see from Figure 3, the introduction of environmental features does not seem to benefit both models. From another side, the classifier did better than the survival analysis model both in terms of AUROC over time and in term of the confidence of its prediction on the long run. As we go further in time, the confidence interval of the survival model widens while the ensemble classifier keeps the same width all along the months following the prediction horizon.

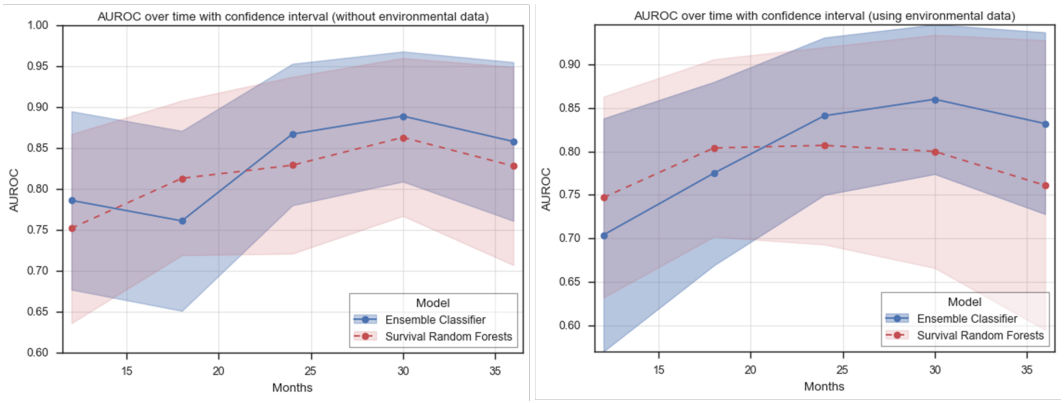

Figure 3: AUROC results for subtask 3-A with 95% confidence interval

For task B, we can clearly see in Figure 4 that the Ensemble classifier model is scoring its best score in terms of AUROC at 36 months. It exceeds its performances in tasks A and C and the reason is obvious: The model was trained and 'hyper-parametrized' on task B data. Noticeably, the width of the confidence interval at 12 months for ensemble classifiers is at its largest. The survival analysis models tend to remarkably outperform ensemble classifiers for the period going from 12 to 30 months when environmental features are introduced. Then, the ensemble classifier reclaims the lead again.

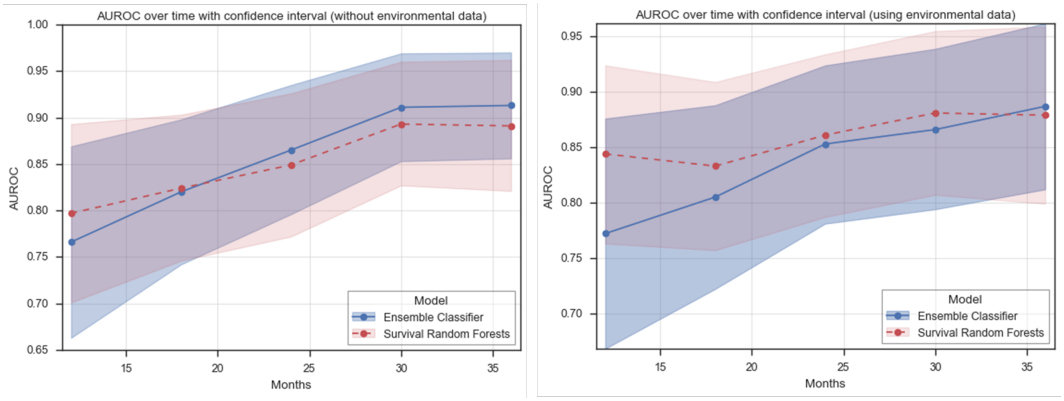

Figure 4: AUROC results for subtask 3-B with 95% confidence interval

Finally, considering the figure 5 depicting AUROC for task C, the first noticeable phenomenon is how unconfident and inaccurate was the ensemble classifier when trying to correctly predict the occurrence of death for a given patient. From that observation, we can assume that knowledge being transferred from task B was mainly related to the occurrence of other events other than death. The co-existence of another event along the 'death' event mitigates the propensity of the classifier to perform badly. The introduction of environmental data to the ensemble classifier goes hand in hand with performance deterioration. The phenomenon was consistently present when assessing the classifier performance through two different ways (C-Index and AUROC).

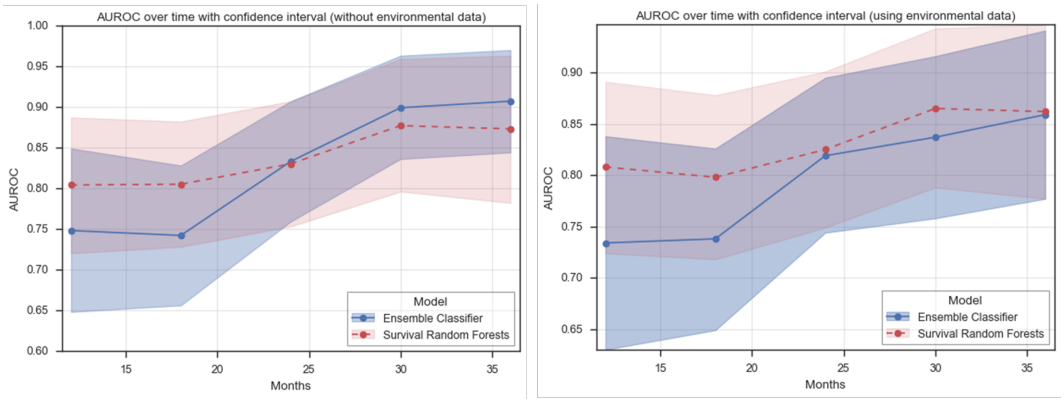

Figure 5: AUROC results for subtask 3-C with 95% confidence interval

All in all, we can conclude that introducing the environmental features to the survival analysis model made it perform better for all the tasks and sometimes with significant gaps with respect to its counterpart trained on static and visits data. The opposite was the case for the ensemble classifier model. However, from an AUROC point of view the ensemble classifier model managed to beat the survival model in the majority of the cases.

Quantitatively speaking, the Ensemble classifier trained on data without environmental features and pre-trained on subtask 3-B data has managed to get the best results for AUROC at prediction horizons beyond 18 months. The assertion holds true for the 3 subtasks: 3-A, 3-B and 3-C. Hence, we can conclude that exploiting knowledge gained from one substask is beneficial for the rest of subtasks when an important proportion of patients is shared between similar prediction tasks. The survival analysis model gets its best results for prediction horizons less than or equal to 18 months. Even though it was hyper-parametrized on subtask 3-A data, the model did not beat pre-trained Ensemble classifiers applied on data for the same task except for 1 case out of 20. Introducing environmental data to prediction models was proven useful only for few cases and the survival model was the main benefiter. However the positive impact of environmental data introduction does not last beyond the 18 months prediction horizon in the best case. Preceding conclusions can be observed from results reported in the following table (table 2).

Table 2
AUROC scores over time of submitted models for subtasks A, B and C

Subtask	Use of Environmental Data	Submission ID	AUROC (12m)	AUROC (18m)	AUROC (24m)	AUROC (30m)	AUROC (36m)
3-A	No	CE_A_base	**0.786 [0.677,0.895]**	0.761 [0.651,0.87]	**0.867 [0.78,0.953]**	**0.889 [0.809,0.968]**	**0.858 [0.761,0.955]**
		SuRF_A_base	0.752 [0.636,0.867]	**0.813 [0.719,0.90]**	0.829 [0.721,0.937]	0.863 [0.767,0.96]	0.828 [0.707,0.949]
	Yes	CE_A_EW6	0.704 [0.57,0.838]	0.775 [0.669,0.88]	0.841 [0.75,0.931]	0.86 [0.774,0.946]	0.832 [0.728,0.937]
		SuRF_A_EW6	0.747 [0.632,0.863]	0.804 [0.702,0.90]	0.807 [0.693,0.92]	0.8 [0.666,0.934]	0.761 [0.595,0.928]
3-B	No	CE_B_base	0.766 [0.663,0.869]	0.82 [0.742,0.89]	**0.865 [0.796,0.935]**	**0.911 [0.853,0.969]**	**0.913 [0.856,0.97]**
		SuRF_B_base	0.797 [0.701,0.893]	0.824 [0.746,0.90]	0.849 [0.772,0.926]	0.893 [0.827,0.96]	0.891 [0.821,0.962]
	Yes	CE_B_EW6	0.772 [0.668,0.876]	0.805 [0.722,0.88]	0.853 [0.781,0.924]	0.866 [0.794,0.939]	0.887 [0.812,0.962]
		SuRF_B_EW6	**0.844 [0.763,0.924]**	**0.833 [0.757,0.90]**	0.861 [0.787,0.934]	0.881 [0.807,0.955]	0.879 [0.799,0.959]
3-C	No	CE_C_base	0.748 [0.648,0.849]	0.742 [0.656,0.82]	**0.833 [0.759,0.907]**	**0.899 [0.836,0.963]**	**0.907 [0.844,0.97]**
		SuRF_C_base	0.804 [0.72,0.887]	**0.805 [0.728,0.88]**	0.83 [0.753,0.907]	0.877 [0.796,0.959]	0.873 [0.782,0.963]
	Yes	CE_C_EW6	0.734 [0.63,0.838]	0.738 [0.649,0.82]	0.819 [0.744,0.895]	0.837 [0.758,0.916]	0.859 [0.777,0.941]
		SuRF_C_EW6	**0.808 [0.724,0.891]**	0.798 [0.718,0.87]	0.825 [0.749,0.901]	0.865 [0.788,0.943]	0.862 [0.777,0.947]

5. Conclusions and Future Work

In this paper summarizing our participation at iDPP 2023 competition, we tried to approach the issue of the impact of environmental data on ALS risk prediction models from a Machine Learning perspective. Our goal can be summarized in the following question: what are the air pollution patterns that characterize groups of ALS patients with fast progression? To answer the question we have begun with stratifying patients relying on the disease progression patterns provided by features extracted from applying staging systems on visits data. The stratification process was performed by a clustering algorithm that does not need a preset number of clusters. Obtained groups of patients were then profiled to determine their common characteristics: clinical, demographic and environmental. Afterwards, a second clustering procedure was carried out to detect clusters of patients with similar exposure concentrations to 3 different air pollutants. Then, the obvious next step was to perform risk prediction on each cluster separately and combine the predictions. However, due to time limitations the step was not reached and predictions were made on the whole population.

Even though we have presented some guidelines to conduct a reliable study based on clustering, our study has some caveats pertaining to the design of the clustering procedure. The lack of available implementations of many necessary tools (clustering algorithms, similarity metrics for mixed-type data, statistical tests ...) hampered our effort to design a robust clustering model. The irregularity of environmental time-series made feature engineering process difficult especially

that ALSFRS-R visits were too variable from one cluster to another. Additionally, the D50 model was not reliable enough to provide us with enough data points that really reflect the disease progress for fast progressing patients. Such factors made diving deeper into studying the causal relationship between air pollutants and disease progression quite difficult and necessitating advanced methods for disease progression modelling and for imputation of missing time-series data. Despite all the encountered obstacles, providing a multimodal real-world dataset for ALS to the research community would make collaborative and reproducible data analysis studies possible which will certainly help the community reach answers to pending questions about ALS in a faster and more certain ways.

For our next steps, we intend to work on improving our patient stratification process by looking for more homogeneous groups. Even though (NO_2) is believed to be a risk factor for ALS, the lack of time-series data for the whole training set pertaining to the air particle prevented us from including it into our air pollution profiling step. We may work on studying pollution patterns for (NO_2) using only patients groups of interest where data are available for more than 50% of the group. Our future work would also include working on improvements of the risk prediction models by studying the possibility of applying one or more models on different clusters and comparing their overall result with models applied on the whole dataset.

Acknowledgments

We would like to thank Guglielmo Faggioli for his valuable assistance and informative answers to our requests throughout the competition.

References

[1] V. Grollemund, P.-F. Pradat, G. Querin, F. Delbot, G. Le Chat, J.-F. Pradat-Peyre, P. Bede, Machine learning in amyotrophic lateral sclerosis: Achievements, pitfalls, and future directions, Front. Neurosci. 13 (2019) 135.

[2] C. Bendotti, V. Bonetto, E. Pupillo, G. Logroscino, A. Al-Chalabi, C. Lunetta, N. Riva, G. Mora, G. Lauria, J. H. Weishaupt, F. Agosta, A. Malaspina, M. Basso, L. Greensmith, L. Van Den Bosch, A. Ratti, M. Corbo, O. Hardiman, A. Chiò, V. Silani, E. Beghi, Focus on the heterogeneity of amyotrophic lateral sclerosis, Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 21 (2020) 485–495. doi:10.1080/21678421.2020.1779298.

[3] M. D. Armstrong, G. Hansen, K. L. Schellenberg, Rural Residence and Diagnostic Delay for Amyotrophic Lateral Sclerosis in Saskatchewan, Canadian Journal of Neurological Sciences 47 (2020) 538–542. doi:10.1017/cjn.2020.38.

[4] H. R. Rashed, M. A. Tork, Diagnostic delay among ALS patients: Egyptian study, Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 21 (2020) 416–419. doi:10.1080/21678421.2020.1763401.

[5] M. A. van Es, O. Hardiman, A. Chio, A. Al-Chalabi, R. J. Pasterkamp, J. H. Veldink, L. H. van den Berg, Amyotrophic lateral sclerosis, The Lancet 390 (2017) 2084–2098. doi:10.1016/S0140-6736(17)31287-4.

[6] C. Karam, J. D. Berry, Heterogeneity, urgency, generalizability, and enrollment: The HUGE balance in ALS trials, Neurology 92 (2019) 215–216. doi:10.1212/WNL.0000000000006837.

[7] G. Faggioli, A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, H. Aidos, R. Bergamaschi, G. Birolo, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García Dominguez, M. Gromicho, E. Longato, S. C. Madeira, U. Manera, G. Silvello, E. Tavazzi, E. Tavazzi, M. Vettoretti, B. Di Camillo, N. Ferro, Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2023, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, A. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2023.

[8] G. Faggioli, A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, H. Aidos, R. Bergamaschi, G. Birolo, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García Dominguez, M. Gromicho, E. Longato, S. C. Madeira, U. Manera, G. Silvello, E. Tavazzi, E. Tavazzi, M. Vettoretti, B. Di Camillo, N. Ferro, Overview of iDPP@CLEF 2023: The Intelligent Disease Progression Prediction Challenge, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), CLEF 2023 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073., 2023.

[9] M. Jankowska-Kieltyka, A. Roman, I. Nalepa, The air we breathe: Air pollution as a prevalent proinflammatory stimulus contributing to neurodegeneration, Front. Cell. Neurosci. 15 (2021) 647643.

[10] L. G. Costa, Traffic-related air pollution and neurodegenerative diseases: Epidemiological and experimental evidence, and potential underlying mechanisms, in: Advances in

Neurotoxicology, Advances in neurotoxicology, Elsevier, 2017, pp. 1–46.
[11] Y. Nunez, A. K. Boehme, J. Goldsmith, M. Li, A. van Donkelaar, M. G. Weisskopf, D. B. Re, R. V. Martin, M.-A. Kioumourtzoglou, PM2.5 composition and disease aggravation in amyotrophic lateral sclerosis: An analysis of long-term exposure to components of fine particulate matter in new york state, Environ. Epidemiol. 6 (2022) e204.
[12] M. Seelen, R. A. Toro Campos, J. H. Veldink, A. E. Visser, G. Hoek, B. Brunekreef, A. J. van der Kooi, M. de Visser, J. Raaphorst, L. H. van den Berg, R. C. H. Vermeulen, Long-term air pollution exposure and amyotrophic lateral sclerosis in netherlands: A population-based case-control study, Environ. Health Perspect. 125 (2017) 097023.
[13] W. Myung, H. Lee, H. Kim, Short-term air pollution exposure and emergency department visits for amyotrophic lateral sclerosis: A time-stratified case-crossover analysis, Environ. Int. 123 (2019) 467–475.
[14] Y. Nunez, A. K. Boehme, M. G. Weisskopf, D. B. Re, A. Navas-Acien, A. van Donkelaar, R. V. Martin, M.-A. Kioumourtzoglou, Fine particle exposure and clinical aggravation in neurodegenerative diseases in new york state, Environ. Health Perspect. 129 (2021) 27003.
[15] M. Povedano, M. Saez, J.-A. Martínez-Matos, M. A. Barceló, Spatial assessment of the association between long-term exposure to environmental factors and the occurrence of amyotrophic lateral sclerosis in catalonia, spain: A population-based nested case-control study, Neuroepidemiology 51 (2018) 33–49.
[16] D. Saucier, P. P. W. Registe, M. Bélanger, C. O'Connell, Urbanization, air pollution, and water pollution: Identification of potential environmental risk factors associated with amyotrophic lateral sclerosis using systematic reviews, Front. Neurol. 14 (2023) 1108383.
[17] E. Tavazzi, E. Longato, M. Vettoretti, H. Aidos, I. Trescato, C. Roversi, A. S. Martins, E. N. Castanho, R. Branco, D. F. Soares, A. Guazzo, G. Birolo, D. Pala, P. Bosoni, A. Chiò, U. Manera, M. de Carvalho, B. Miranda, M. Gromicho, I. Alves, R. Bellazzi, A. Dagliati, P. Fariselli, S. C. Madeira, B. Di Camillo, Artificial intelligence and statistical methods for stratification and prediction of progression in amyotrophic lateral sclerosis: A systematic review, Artif. Intell. Med. 142 (2023) 102588.
[18] C. Hennig, What are the true clusters?, Pattern Recognition Letters 64 (2015) 53–62. doi:10.1016/j.patrec.2015.04.009.
[19] A. Zimek, Clustering High-Dimensional Data, in: Data Clustering, 1 ed., Chapman and Hall/CRC, 2014, pp. 201–230. doi:10.1201/9781315373515-9.
[20] C. C. Aggarwal, A. Hinneburg, D. A. Keim, On the Surprising Behavior of Distance Metrics in High Dimensional Space, in: G. Goos, J. Hartmanis, J. van Leeuwen, J. Van den Bussche, V. Vianu (Eds.), Database Theory — ICDT 2001, volume 1973, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 420–434. doi:10.1007/3-540-44503-X_27.
[21] A. Adolfsson, M. Ackerman, N. C. Brownstein, To cluster, or not to cluster: An analysis of clusterability methods, Pattern Recognit. 88 (2019) 13–26.
[22] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognit. 46 (2013) 243–256.
[23] Theme 10 - DISEASE STRATIFICATION AND PHENOTYPING OF PATIENTS, Amyotroph. Lateral Scler. Frontotemporal Degener. 22 (2021) 174–185.
[24] R. Kueffner, N. Zach, M. Bronfeld, R. Norel, N. Atassi, V. Balagurusamy, B. Di Camillo, A. Chio, M. Cudkowicz, D. Dillenberger, J. Garcia-Garcia, O. Hardiman, B. Hoff, J. Knight,

M. L. Leitner, G. Li, L. Mangravite, T. Norman, L. Wang, ALS Stratification Consortium, J. Xiao, W.-C. Fang, J. Peng, C. Yang, H.-J. Chang, G. Stolovitzky, Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach, Sci. Rep. 9 (2019) 690.

[25] M. Tang, C. Gao, S. A. Goutman, A. Kalinin, B. Mukherjee, Y. Guan, I. D. Dinov, Model-based and model-free techniques for amyotrophic lateral sclerosis diagnostic prediction and patient clustering, Neuroinformatics 17 (2019) 407–421.

[26] J. Wu, Advances in K-means clustering, Springer theses, 2012 ed., Springer, Berlin, Germany, 2012.

[27] J. C. Roche, R. Rojas-Garcia, K. M. Scott, W. Scotton, C. E. Ellis, R. Burman, L. Wijesekera, M. R. Turner, P. N. Leigh, C. E. Shaw, A. Al-Chalabi, A proposed staging system for amyotrophic lateral sclerosis, Brain 135 (2012) 847–852.

[28] N. J. Thakore, B. R. Lapin, T. G. Kinzy, E. P. Pioro, Deconstructing progression of amyotrophic lateral sclerosis in stages: a markov modeling approach, Amyotroph. Lateral Scler. Frontotemporal Degener. 19 (2018) 483–494.

[29] K. Poesen, M. De Schaepdryver, B. Stubendorff, B. Gille, P. Muckova, S. Wendler, T. Prell, T. M. Ringer, H. Rhode, O. Stevens, K. G. Claeys, G. Couwelier, A. D'Hondt, N. Lamaire, P. Tilkin, D. Van Reijen, S. Gourmaud, N. Fedtke, B. Heiling, M. Rumpel, A. Rödiger, A. Gunkel, O. W. Witte, C. Paquet, R. Vandenberghe, J. Grosskreutz, P. Van Damme, Neurofilament markers for ALS correlate with extent of upper and lower motor neuron disease, Neurology 88 (2017) 2302–2309.

[30] A. Shaabi, Modeling amyotrophic lateral sclerosis progression: Logic in the logit, Cureus 14 (2022) e24887.

[31] R. Balendra, A. Jones, N. Jivraj, C. Knights, C. M. Ellis, R. Burman, M. R. Turner, P. N. Leigh, C. E. Shaw, A. Al-Chalabi, Estimating clinical stage of amyotrophic lateral sclerosis from the ALS functional rating scale, Amyotroph. Lateral Scler. Frontotemporal Degener. 15 (2014) 279–284.

[32] A. Genge, A. Chio, The future of ALS diagnosis and staging: where do we go from here?, Amyotroph. Lateral Scler. Frontotemporal Degener. 24 (2023) 165–174.

[33] A. Chiò, E. R. Hammond, G. Mora, V. Bonito, G. Filippini, Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis, J. Neurol. Neurosurg. Psychiatry 86 (2015) 38–44.

[34] H.-J. Westeneng, T. P. A. Debray, A. E. Visser, R. P. A. van Eijk, J. P. K. Rooney, A. Calvo, S. Martin, C. J. McDermott, A. G. Thompson, S. Pinto, X. Kobeleva, A. Rosenbohm, B. Stubendorff, H. Sommer, B. M. Middelkoop, A. M. Dekker, J. J. F. A. van Vugt, W. van Rheenen, A. Vajda, M. Heverin, M. Kazoka, H. Hollinger, M. Gromicho, S. Körner, T. M. Ringer, A. Rödiger, A. Gunkel, C. E. Shaw, A. L. Bredenoord, M. A. van Es, P. Corcia, P. Couratier, M. Weber, J. Grosskreutz, A. C. Ludolph, S. Petri, M. de Carvalho, P. Van Damme, K. Talbot, M. R. Turner, P. J. Shaw, A. Al-Chalabi, A. Chiò, O. Hardiman, K. G. M. Moons, J. H. Veldink, L. H. van den Berg, Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model, Lancet Neurol. 17 (2018) 423–433.

[35] T. Prell, N. Gaur, R. Steinbach, O. W. Witte, J. Grosskreutz, Modelling disease course in amyotrophic lateral sclerosis: pseudo-longitudinal insights from cross-sectional health-related quality of life data, Health Qual. Life Outcomes 18 (2020) 117.

[36] R. J. G. B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical

density estimates, in: Advances in Knowledge Discovery and Data Mining, Lecture notes in computer science, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 160–172.
[37] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, Efficient and Robust Automated Machine Learning, in: Advances in Neural Information Processing Systems, volume 28, 2015.
[38] Z. Qian, B.-C. Cebere, M. van der Schaar, Synthcity: facilitating innovative use cases of synthetic data in different data modalities (2023).

A. Detailed Profiles of Clustered Patients

The present appendix provides a detailed data profiling of clusters of patients resulting from the clustering process described in the methodology section. Tables A.1 and A.2 provide aggregated characteristics (Mean and Coefficient of Variation) of features resulting from 3 ALS staging systems (MiToS, Fine-Till-Nine and D50). The tables also highlight differences between the totality of patients belonging to a cluster and the subpopulation (derived from each cluster) having environmental data. In the aforementioned tables, 'NAD' and 'NAFG' refer to *No Affected Domains* and *No Affected Functional Groups* respectively.

For table A.3, we have presented characteristics of patients belonging to each cluster and having environmental data. The described patients are originating from the pool of patients used to build training sets for the iDPP subtasks 3-A, 3-B and 3-C. We have only presented results pertaining to features that were deemed to be important for clinicians to differentiate between groups of ALS patients. Patients' features that are summarized by the table pertain to static and environmental data modalities of the iDPP 2023 dataset. The different features were further grouped into different sets namely: Onset Site, Affected Limbs, Medical History, Presence of ALS-charactersitic gene mutations as well as the group of environmental characteristics. For the continuous features, we have provided the feature mean value for each cluster as well as the coefficient of variation put between brackets. For the rest of features we have provided percentages of selected values with respect to the total number of patients for each cluster. For the 'Affected Limbs' group, we have additionally computed percentages of groups of patients having a given value with respect to the subgroup of patients belonging to each cluster whose ALS onset site was the limbs. The additional percentage was presented as a ratio and placed between brackets.

In order to improve the readability of Tables A.1, A.2 and A.3, we have added tables mapping categories to codes that have been used during our data analysis. Tables A.4, A.5 and A.6 provide mappings for Fine-Till-9 affected groups, patient occupation and patient disease history. Using shades of different colors to highlight important pieces of information about clusters was another method to facilitate the readability of different patients profiles.

Table A.1

Clusters **(All Patients | Patients with ENV data);**

Cluster Label	0		1		2		3		4		5		6		7 (1 pat(ENV)=5.3%)		8 (1 pat(ENV)=5.9%)		9		10 (1 pat(ENV)=8.3%)	
Number of Patients	**124**	**13**	**120**	**32**	**124**	**35**	**103**	**28**	**165**	**43**	**155**	**41**	**108**	**39**	**87**	**19**	**74**	**17**	**82**	**9**	**95**	**12**
Initial MiToS Aff. Grps	**NAD**:99.2%	100%	100%		100%		100%				99.4%	100%	100%		96.6%	94.7%	70.3%	64.7%	NAD: 100%		NAD: 70.5%	75%
																	Movement:29.7%	35.3%			Movement: 18.9%	16.7%
																					Breathing: 4.2%	8.3%
Final MiToS Aff. Grps	Absent: 100%		99.2%	96.9%			93.2%	92.9%	100%						NAD: 73.6%	63.2%	Movement: 86.5%	88.2%	Absent: 100%			
															Movement: 12.6%	26.3%	Breathing: 6.8%	5.9%				
															Breathing:	5.3%						
															Swallowing: 9.2%	5.3%	Swallowing: 4.1%					
																	Movement & Breathing: 2.7%	5.9%				
Initial Fine_9 Aff. Grps	Gross Motor: 52.4% Fine Motor: 25.8% Bulbar: 21%	0: 53.8% 12: 46.2%	NAFG: 93.3%	90.6%	Fine Motor & Gross Motor:46.8% NAFG:36.3% Bulbar & Fine Motor:8.9%	9: 45.7% 14: 34.3% 5: 8.6% 1: 8.6%	**Gross Motor: 100%**		**NAFG:100%**		Gross Motor: 35.5% Bulbar: 25.8% Fine Motor: 19.4% NAFG: 15.5%	12: 29.3% 14: 24.4% 8: 22% 0: 19.5%	Bulbar: 89.8%	92.3%	Fine Motor: 83.9%	84.2%	Fine Motor & Gross Motor: 82.4%		NAFG: 100%		Fine Motor & Gross Motor: 43.2% Bulbar & Fine Motor & Gross Motor: 24.2% Bulbar & Gross Motor: 8.4% Bulbar & Fine Motor: 8.4% Fine Motor & Gross Motor & Respiratory: 6.3% Bulbar & Fine Motor & Gross Motor & Respiratory: 4.2% Bulbar & Gross Motor & Respiratory: 3.2%	2: 33.3% 9: 33.3% 6: 8.3% 4: 8.3% 10: 8.3% 13: 8.3%
													Fine Motor & Gross Motor: 3.7%		NAFG: 5.7%	10.5%	9.5%	11.8%				
													Bulbar & Fine Motor: 3.7%	5.1%	Bulbar: 5.7%	5.3%						
Final Fine_9 Aff. Grps	Absent: 100%		**NAFG: 100%**		Fine Motor & Gross Motor:66.9%	62.9%			Bulbar: 35.2% Gross Motor: 32.1% Fine Motor: 31.5%	12: 44.2% 0: 34.9% 8: 20.9%	Fine Motor & Gross Motor: 60.6% Bulbar & Fine Motor: 17.4% Bulbar & Gross Motor: 15.5%	9: 58.5% 5: 26.8% 7: 4.9% 13: 4.9%	Bulbar: 93.5%	92.3%	Fine Motor: 86.2%	89.5%	Fine Motor & Gross Motor: 86.5%	88.2%	Absent: 100%			
					Bulbar & Gross Motor:18.5%	22.9%							Fine Motor: 4.6%	2.6%	Bulbar: 9.2%	5.3%	Fine Motor & Respiratory: 4.1%	5.9%				
					Bulbar & Fine Motor:9.7%	8.6%							Gross Motor: 1.9%	5.1%	Respiratory: 4.6%	5.3%	Bulbar & Fine Motor: 2.7%					
																	Bulbar & Respiratory: 2.7%	5.9%				
Last_MiToS_Stage	0: 99.2%	100%	99.2%	96.6%	100%		93.2%	92.9%	100%						0: 73.6% 1: 26.4%	0:63.2% 1: 36.8%	1: 97.3%	94.1%	0: 100%		0: 70.5%	75%
																					1: 24.2%	25%
																					2: 5.3%	
Last_Fine_9_Stage	1: 100%		**0: 100%**		2: 100%		1: 100%		1: 100%		2: 98.7%	97.6%	**1: 100%**		**1: 100%**		2: 98.6%	100%	0: 100%		2: 61.1% 3: 34.7% 4: 4.2%	3: 58.3% 2: 41.7%
											4: 1.3%	2.4%										
MiToS_Stage0_Norma_Dur	1: 99.2%	100%	1:99.2%	96.9%	100%		0.99 (0.041)	0.99 (0.017)	1:100%		1: 99.4%	100%	1: 100%		0.97 (0.105)	0.94 (0.12)	0.78 (0.185)	0.73 (0.23)	1: 100%		0.92 (0.17)	0.93 (0.14)
MiToS_Stage1_Norma_Dur	0: 99.2%	100%	0: 99.2%	96.9%	100%		0.007 (5.8)	0.004 (4.3)	0 :100%		0: 99.4%	100%	0 :100%		0.03 (3.16)	0.06 (1.92)	0.22 (0.676)	0.25 (0.71)	0: 100%		0.07 (2.06)	0.07 (1.87)
MiToS_Stage2_Norma_Dur	**0: 100%**																0: 97.3%	94.1%			0.01 (6.07)	0: 100%
MiToS_Stage3_Norma_Dur																	0: 100%				0: 100%	
MiToS_Stage4_Norma_Dur																						
Fine_9_Stage0_Norma_Dur	0.63 (0.254)	0.62 (0.245)	0.98 (0.073)	0.98 (0.07)	0.71 (0.2)	0.73 (0.19)	0.64 (0.2)	0.65 (0.21)	0.79 (0.149)	0.77 (0.17)	0.65 (0.22)	0.68 (0.2)	0.60 (0.201)	0.61 (0.17)	0.61 (0.251)	0.63 (0.22)	0.61 (0.203)	0.57 (0.19)	1: 100%		0.71 (0.25)	0.68 (0.26)
Fine_9_Stage1_Norma_Dur	0.37 (0.433)	0.38 (0.402)	0.01 (4.202)	0.01 (4.3)	0: 98%	0: 94.3%	0.35 (0.35)	0.35 (0.4)	0.21 (0.577)	0.23 (0.57)	0.21 (0.601)	0.19 (0.61)	0.38 (0.367)	0.37 (0.33)	0.38 (0.431)	0.36 (0.42)	0: 98.6%	100%	0: 100%		0: 100%	
Fine_9_Stage2_Norma_Dur	0: 100%		0: 100%		0.28 (0.49)	0.27 (0.52)	**0: 100%**				0.13 (0.76)	0.12 (0.71)	0.02 (3.383)	0.017 (4.18)	0: 95.4%	0: 94.7%	0.39 (0.324)	0.43 (0.24)			0.18 (1.083)	0.15 (1.45)
Fine_9_Stage3_Norma_Dur			0: 99.2%	0: 96.9%	0: 99.2%	0: 100%					0.005 (6.042)	0.01 (4.471)	0: 100%								0.1 (1.803)	0.17 (1.17)
Fine_9_Stage4_Norma_Dur			0:100%		0: 100%						0: 98.7%	0: 97.6%	0: 100%		0: 98.9%	0: 100%					0: 95.8%	100%
Early_ALSFRS-R_slope (Mean (CV))	-0.66 (1.084)	-0.65 (-0.71)	**-0.28 (-0.983)**	**-0.28 (-1.19)**	-0.67 (-0.69)	-0.61 (-0.63)	-0.54 (-0.89)	-0.52 (-0.67)	-0.33 (-0.73)	-0.32 (-0.57)	-0.55 (-0.753)	-0.49 (-0.88)	-0.59 (-0.682)	-0.58 (-0.57)	-0.57 (-0.833)	-0.52 (-0.72)	**-1.26 (-0.65)**	**-1.42 (-0.6)**	-0.43 (-1.4)	-0.22 (-0.4)	-1.24 (-1.813)	-1.11 (-0.68)
Late_ALSFRS-R_slope	Absent: 100%		**-0.17 (-2.78)**	**-0.1 (-6.552)**	-0.84 (-0.874)	-0.77 (-0.75)	-0.47 (-1.077)	-0.47 (-1.07)	-0.78 (-0.823)	-0.93 (-0.90)	-1.29 (-0.614)	-1.27 (-0.68)	-0.43 (-1.51)	-0.35 (-1.41)	-0.56 (-0.985)	-0.67 (-0.93)	**-1.50 (-0.659)**	**-1.53 (-0.37)**	Absent: 100%			
Number of Visits	1: 100%		2:52.5% 3: 35% 4: 10%	3:46.9% 2:34.4% 4:18.8%	3: 45.2% 2: 36.3% 4: 14.5%	3:48.6% 2:37.1% 4:11.4%	2:50.5% 3: 32% 4:12.6%	2:46.4% 3:42.9% 4:7.1% 7: 3.6%	3:37.6% 2:37.6% 4: 17% 5: 6.1%	3:39.5% 2:37.2% 4:16.3% 5: 7%	3:46.5% 2:32.9% 4:14.8%	3:43.9% 2:29.3% 4:17.1%	2:45.4% 3:36.1% 4:16.7%	2: 41% 3:33.3% 4:23.1%	2: 46% 3: 40.2% 4: 11.5%	2:57.9% 3:42.1%	2: 44.6% 3: 37.8% 4: 14.9%	3:52.9% 2:29.4% 4:17.6%	1: 100%			
Diagnosis Delay	12.63 (0.766)	12.91 (1.023)	12.91 (1.188)	13.42 (0.9)	12.44 (0.88)	14.18 (0.93)	11.71 (0.51)	12 (0.54)	7.91 (0.639)	7.40 (0.62)	12.23 (0.793)	13.55 (0.90)	9.26 (0.474)	9 (0.45)	9.55 (0.541)	9.58 (0.59)	9.11 (0.5)	7.1 (0.56)	10.61 (0.9)	10.83 (0.6)	17.23 (0.99)	17.92 (1.18)
D50_stage	Absent: 100%		1:72.5% 2:15.8% 3:11.7%	1:68.8% 2:15.6% 3:15.6%	2: 83.9% 3: 12.9% 1: 3.2%	2:80% 3:17.1% 1: 2.9%	2:75.7% 1:24.3%	2: 75% 1: 25%	2: 63% 1: 37%	2:65.1% 1:34.9%	2:85.8% 3:11.6%	2:82.9% 3:12.2%	2:64.8% 1:35.2%	2:64.1% 1:35.9%	2: 67.8% 1: 32.2%	2:84.2% 1:15.8%	2: 91.9% 3: 8.1%	94.1% 5.9%	Absent: 100%			

Clusters **(All Patients | Patients with ENV data);**

Table A.2

11 (1 pat(ENV)=4%)		12 (1 pat(ENV)=2%)		13 (1 pat(ENV)=3.2%)		14 (1 pat(ENV)=2.6%)		
77	**25**	**158**	**49**	**94**		**115**	**38**	Number of Patients
NAD: 100%		NAD: 69.6%	81.6%	93.6%	93.5%	55.7%	47.4%	Initial MiToS Aff. Grps
		Movement: 22.2%	10%	5.3%	6.5%	40%	52.6%	
		Breathing: 2.5%	2%	1.1%	0%	4.3%	0%	
		Movement: 28.5%	26.5%	68.1%	80.6%	60%	68.5%	Final MiToS Aff. Grps
		Breathing: 20.3%	34.7%	19.1%	12.9%	22.6%	18.4%	
		Movement & Breathing: 19%	16.3%	4.3%	6.5%	8.7%	7.9%	
		Movement & Communicating: 10.8%	8.2%					
		Movement & Swallowing: 7%	8.2%					
		Movement & Swallowing & Communicating: 5.1%				1.7%	2.6%	
		Swallowing: 3.2%		7.4%	6.5%	3.5%	0%	
Fine Motor & Gross Motor: 24.7% Bulbar: 2.4% NAFG:22.1% Fine Motor: 10.4% Bulbar & Fine Motor: 6.5% Gross Motor: 6.5% Bulbar & Gross Motor: 6.5%	14: 40% 9: 20% 0: 16% 8: 12% 5: 8%	Bulbar & Fine Motor & Gross Motor: 32.9% Fine Motor & Gross Motor: 13.3% NAFG: 12.7% Fine Motor & Gross Motor & Respiratory: 10.1% Bulbar & Fine Motor & Gross Motor & Respiratory: 8.2% Bulbar: 5.7% Bulbar & Fine Motor: 5.1% Gross Motor: 3.8%	2: 24.5% 14: 18.4% 9: 16.3% 10: 16.3% 0: 6.1%	Gross Motor: 48.9% Fine Motor: 29.8% Bulbar: 10.6% NAFG: 8.5%	61.3% 25.8% 3.2% 9.7%	Fine Motor & Gross Motor: 66.1% Fine Motor: 7% Bulbar: 6.1% NAFG: 3.5% Bulbar & Gross Motor: 3.5% Fine Motor & Respiratory: 3.5% Gross Motor: 2.6% Bulbar and Respiratory: 2.6%	78.9% 10.5% 5.3% 0% 2.6% 2.6%	Initial Fine_9 Aff. Grps
Bulbar & Fine Motor & Gross Motor: 83.1%	72%	Bulbar & Fine Motor & Gross Motor & Respiratory: 52.5%	51%	Fine Motor & Gross Motor: 53.2% Bulbar & Fine Motor & Gross Motor: 9.6%	61.3% 12.9%	Bulbar & Fine Motor & Gross Motor: 33.9% Fine Motor & Gross Motor & Respiratory: 33.9% Fine Motor & Gross Motor: 18.3% Bulbar & Fine Motor & Respiratory: 6.1%	28.9% 39.5% 18.4% 10.5%	Final Fine_9 Aff. Grps
		Bulbar & Fine Motor & Gross Motor: 30.4%	26.5%	Fine Motor & Gross Motor & Respiratory: 8.5% Bulbar & Respiratory: 7.4%	6.5% 6.5%	Bulbar & Fine Motor & Gross Motor & Respiratory: 3.5% Bulbar & Gross Motor & Respiratory: 2.6%	2.6% 0%	
Fine Motor & Gross Motor & Respiratory: 10.4%	16%	Fine Motor & Gross Motor & Respiratory: 15.2%	20.4%	Gross Motor & Respiratory: 6.4% Bulbar & Gross Motor: 4.3% Bulbar & Fine Motor & Gross Motor & Respiratory: 4.3% Bulbar & Fine Motor: 3.2%				
Bulbar & Gross Motor & Respiratory: 5.2%	12%			Fine Motor & Respiratory: 3.2%	6.5%			
0: 100%		1: 51.9%	61.2%	1: 94.7%	93.5%	1: 87%	86.8%	Last_MiToS_Stage
		2: 37.3%	32.7%	2: 5.3%	6.5%	2: 10.4%	7.9%	
		3: 8.9%	6.1%			3: 1.7%	2.6%	
3: 100%		**4: 52.5%**	**51%**	2: 77.7%	77.4%	3: 76.5%	78.9%	Last_Fine_9_Stage
		3: 47.5%	49%	3: 18.1%	19.4%	2: 20	18.4%	
				4: 4.3%	3.2%	4: 3.5%	2.6%	
1: 100%		0.80 (0.16)	0.81 (0.16)	0.83 (0.137)	0.85 (0.12)	0.8 (0.17)	0.78 (0.19)	MiToS_Stage0_Norma_Dur
0: 100%		0.13 (0.97)	0.14 (0.83)	0.16 (0.67)	0.15 (0.64)	0.18 (0.72)	0.21 (0.68)	MiToS_Stage1_Norma_Dur
		0.05 (1.76)	0.05 (1.79)	0.01 (5.485)	0.003 (4.61)	0.01 (4.18)	0.004 (4.54)	MiToS_Stage2_Norma_Dur
		0.01 (3.49)	0.01 (4.06)	0: 100%		98.3%	97.4%	MiToS_Stage3_Norma_Dur
		0: 98.1%	100%					MiToS_Stage4_Norma_Dur
0.64 (0.19)	0.65 (0.17)	0.65 (0.23)	0.64 (0.26)	0.61 (0.24)	0.63 (0.27)	0.69 (0.242)	0.68 (0.22)	Fine_9_Stage0_Norma_Dur
0.1 (1.16)	0.1 (1.17)	0.04 (2.658)	0.03 (2.86)	0.2 (0.59)	0.2 (0.57)	0.05 (2.35)	0.05 (2.28)	Fine_9_Stage1_Norma_Dur
0.11 (1.13)	0.12 (1.1)	0.06 (2.019)	0.07 (1.914)	0.16 (0.7)	0.14 (0.72)	0.15 (0.83)	0.16 (0.76)	Fine_9_Stage2_Norma_Dur
0.15 (0.617)	0.14 (0.53)	0.18 (1)	0.19 (0.95)	0.03 (2.45)	0.03 (2.4)	0.11 (0.94)	0.10 (0.76)	Fine_9_Stage3_Norma_Dur
0: 98.7%	100%	0.08 (1.472)	0.06 (1.64)	0: 95.7%	96.8%	0.004 (6.134)	0.005 (6.16)	Fine_9_Stage4_Norma_Dur
-0.81 (-0.823)	**-0.74 (-0.69)**	**-1.4 (-0.733)**	**-1.34 (-0.73)**	**-0.75 (-0.76)**	**-0.81 (-0.9)**	-1.07 (-1)	-1.17 (-0.94)	Early_ALSFRS-R_slope (Mean (CV))
-1.89 (-0.506)	**-2.2 (-0.49)**	**-2.58 (-0.647)**	**-2.72 (-0.6)**	**-2.22 (-0.445)**	**-2.1 (-0.49)**	-2.04 (-0.67)	-1.88 (-0.64)	Late_ALSFRS-R_slope
2: 41.6%	40%	3: 44.3%	44.9%	47.9%	51.6%	2: 49.6%	55.3%	Number of Visits
3: 36.4%	28%	2: 38.6%	32.7%	4: 23.4%	16.1%	3: 36.5%	28.9%	
4: 18.2%	24%	4: 12.7%	14.3%	2: 23.4%	32.3%	4: 11.3%	10.5%	
5: 3.9%	8%	5: 4.4%	8.2%	5: 4.3%		5: 2.6%	5.3%	
9.37 (0.528)	8.31 (0.56)	11.97 (1)	9.82 (0.66)	10 (0.768)	11.54 (0.86)	16.3 (0.818)	16.79 (0.74)	Diagnosis Delay
2: 100%		3: 67.1%	67.3%	2: 84%	74.2%	3: 55.7%	55.3%	D50_stage
		2: 32.9%	32.7%	3: 16%	25.8%	2: 44.3%	44.7%	

Table A.3

Clusters **(Patients with ENV data);**

		0 (1 pat=7.7%)	1 (1 pat=3.1%)	2 (1 pat=2.9%)	3 (1 pat=3.6%)	4 (1 pat=2.3%)	5 (1 pat=2.4%)	6 (1 pat=2.6%)	7 (1 pat=5.3%)	8 (1 pat=5.9%)	9 (1 pat=11.1%)	10 (1 pat=8.3%)	11 (1 pat=4%)	12 (1 pat=2%)	13 (1 pat=3.2%)	14 (1 pat=2.6%)	
Number of Patients		13	32	35	28	43	41	39	19	17	9	12	25	49	31	38	$\sum = 431\ (92\%)$
Alive	True	23.1%	53.1%	34.3%	42.9%	27.9%	41.5%	12.8%	15.8%	29.4%	33.3%	16.7%	24%	14.3%	22.6%	26.3%	
Sex	Male	23.1%	68.8%	51.4%	42.9%	46.5%	56.1%	43.6%	68.4%	64.7%	33.3%	41.7%	44%	46.9%	38.7%	55.3%	
Age_onset (Mean (CV))		65.13 (0.17)	62.98 (0.13)	63.85 (0.21)	63.68 (0.15)	63.48 (0.17)	67.21 (0.13)	66.14 (0.14)	67.52 (0.13)	64.63 (0.18)	**58.29 (0.2)**	**69.4 (0.17)**	67.7% (0.12)	65.39 (0.16)	67.9 (0.15)	66.2 (0.21)	
prevalentLMN	True	30.8%	28.1%	48.6%	67.9%	27.9%	41.5%	15.4%	42.1%	58.8%	22.2%	33.3%	12%	36.7%	51.6%	42.1%	
	Absent	7.7%	3.1%	2.9%	3.6%	0%	4.9%	15.4%	5.3%	0%	22.2%	16.7%	0%	10.2%	9.7%	15.8%	
prevalentUMN	True	38.5%	9.4%	17.1%	17.9%	16.3%	12.2%	28.2%	0%	11.8%	22.2%	41.7%	28%	18.4%	12.9%	18.4%	
	Absent	7.7%	3.1%	2.9%	3.6%	0%	4.9%	15.4%	5.3%	0%	22.2%	16.7%	0%	10.2%	9.7%	15.8%	
mixedMN	True	23.1%	59.4%	31.4%	10.7%	55.8%	41.5%	41%	52.6%	29.4%	33.3%	8.3%	60%	34.7%	25.8%	23.7%	
	Absent	7.7%	3.1%	2.9%	3.6%	0%	4.9%	15.4%	5.3%	0%	22.2%	16.7%	0%	10.2%	9.7%	15.8%	
Onset_bulbar	True	53.8%	25%	20%	3.6%	39.5%	31.7%	94.9%	10.5%	5.9%	22.2%	41.7%	56%	42.9%	9.7%	10.5%	**Onset Site**
Onset_axial	True	7.7%	3.1%	0%	0%	0%	0%	0%	0%	0%	11.1%	8.3%	0%	8.2%	0%	2.6%	
Onset_generalized	True	0%	0%	2.9%	0%	0%	0%	2.6%	5.3%	0%	0%	0%	0%	0%	0%	2.6%	
Onset_Limbs	True	**38.5%**	**71.9%**	**80%**	**96.4%**	**60.5%**	**68.3%**	**5.1%**	**89.5%**	**94.1%**	**66.7%**	**50%**	**44%**	**49%**	**90.3%**	**86.8%**	
Onset_limb_UpLo	Lower	38.5% (1)	28.1% (0.39)	37.1% (0.46)	**85.7% (0.89)**	34.9% (0.58)	43.9% (0.64)	5.1% (1)	5.3% (0.06)	41.2% (0.44)	44.4% (0.67)	33.3% (0.67)	8% (0.18)	26.5% (0.54)	58.1% (0.64)	52.6% (0.61)	**Affected Limbs**
	Upper		43.8% (0.61)	40% (0.5)	10.7% (0.11)	25.6% (0.42)	24.4% (0.36)		84.2% **(0.94)**	52.9% (0.56)	22.2% (0.33)	16.7% (0.33)	32% (0.72)	20.4% (0.42)	32.3% (0.36)	31.6% (0.36)	
	Up_Low												4% (0.1)	2% (0.04)			
	Limb			2.9% (0.04)												2.6% (0.03)	
Onset_limb_RiLE	Right	7.7% (0.2)	25% (0.35)	28.6% (0.36)	50% (0.52)	32.6% (0.54)	22% (0.32)		52.6% (0.59)	52.9% (0.56)	33.3% (0.5)	33.3% (0.67)	32% (0.73)	20.4% (0.42)	41.9% (0.46)	44.7% (0.51)	
	Left	15.4% (0.4)	34.4% (0.48)	28.6% (0.36)	35.7% (0.37)	18.6% (0.31)	24.4% (0.36)	5.1% (1)	21.1% (0.24)	23.5% (0.25)	22.2% (0.33)		4% (0.09)	18.4% (0.38)	29% (0.32)	28.9% (0.33)	
	Right_Left	15.4% (0.4)	12.5% (0.18)	20% (0.25)	10.7% (0.11)	9.3% (0.15)	22% (0.32)		15.8% (0.18)	17.6% (0.19)	11.1% (0.17)	16.7% (0.33)	8% (0.18)	8.2% (0.17)	19.4% (0.21)	5.3% (0.06)	
	Limb			2.9% (0.04)												2.6% (0.03)	
	Other													2% (0.04)		5.3% (0.03)	
Onset_limb_DisPro	Distal	23.1% (0.6)	59.4% (0.83)	65.7% (0.82)	78.6% (0.82)	46.5% (0.77)	53.7% (0.79)		68.4% (0.76)	70.6% (0.75)	55.6% (0.83)	25% (0.5)	36% (0.82)	36.7% (0.75)	54.8% (0.61)	52.6% (0.61)	
	Proximal	7.7% (0.2)	6.2% (0.09)	5.7% (0.07)	10.7% (0.11)	14% (0.23)	7.3% (0.11)	5.1% (1)	21.1% (0.24)	11.8% (0.13)		8.3% (0.17)		6.1% (0.12)	22.6% (0.25)	5.3% (0.06)	
	Dis_Pro	7.7% (0.2)	3.1% (0.04)	2.9% (0.04)	3.6% (0.04)		4.9% (0.07)			11.8% (0.13)		8.3% (0.17)	8% (0.18)	2% (0.04)	12.9% (0.14)	10.5% (0.12)	
	Limb			2.9% (0.04)												2.6% (0.03)	
	Other		3.1% (0.04)	2.9% (0.04)	3.6% (0.04)		2.4% (0.04)				11.1% (0.17)	8.3% (0.17)		4.1% (0.08)		15.8% (0.18)	
Occupation	Absent	46.2%	40.6%	28.6%	25%	23.3%	29.3%	30.8%	21.1%	17.6%	**55.6%**	41.7%	28%	28.6%	22.6%	39.5%	
	Most Freq.	**40: 15.4%**	**36: 9.4%**	**18: 8.6%**	**27: 10.7%** 15: 10.7%	**18: 9.3%** **5: 9.3%**	15: 12.2% **22: 9.8%**	**36: 10.3%** **14: 7.7%**	**5: 15.8%** **40: 15.8%** 25: 10.5% **22: 10.5%**	**27: 17.6%** 15: 11.8% **5: 11.8%**	0: 11.1% 37: 11.1% **40: 11.1%** 15: 11.1%	30: 8.3% **14: 8.3%** 29: 8.3% **39: 8.3%**	**18: 8%** **39: 8%** 15: 8% 4: 8%	7: 8.2% 15: 6.1% 23: 6.1% 3: 6.1%	15: 9.7% **5: 9.7%** 10: 6.5% 32: 6.5%	**40: 10.5%** **22: 7.9%**	
Number_of_Diseases	Absent	23.1%	34.4%	**40%**	32.1%	37.2%	19.5%	28.2%	15.8%	29.4%	33.3%	25%	24%	20.4%	29%	31.6%	**Medical History**
	Most Freq.	**2: 30.8%** 1: 23.1%	1: 46.9% 4: 6.2%	1: 28.6% 2: 25.7%	**2: 39.3%** 1: 21.4%	1: 41.9% 3: 11.6%	1: 36.6% 3: 22% 2: 19.5%	1: 33.3% 3: 20.5% 2: 12.8%	1: 52.6% 2: 26.3%	**2: 41.2%** 1: 29.4%	1: 33.3% 2: 22.2% 3: 11.1%	1: 33.3% **3: 25%** 2: 16.7%	**2: 32%** 1: 24% 3: 12%	**2: 36.7%** 1: 30.6% 3: 10.2%	1: 22.6% 3: 19.4% 2: 16.1% 4: 12.9%	2: 28.9% 1: 21.1% 3: 15.8%	
Disease_History	Absent	23.1%	34.4%	**40%**	32.1%	37.2%	19.5%	28.2%	15.8%	29.4%	33.3%	25%	24%	20.4%	29%	31.6%	
	Most Freq.	22: 15.4% **61: 15.4%**	22: 21.9% **17: 18.8%**	**22: 22.9%** **47: 8.6%**	**47: 21.4%** **22: 17.9%**	**22: 25.6%** **17: 7%**	**22: 22%** **47: 9.8%**	22: 23.1% **50: 7.7%**	22: 26.3% **17: 15.8%** **66: 10.5%**	47: 17.6% **17: 11.8%** **69: 11.8%** **27: 11.8%**	**22: 22.2%** **47: 11.1%** **56: 11.1%** **61: 11.1%**	**22: 16.7%** **61: 8.3%** **17: 8.3%** **47: 8.3%**	**47: 16%** **22: 16%** 32: 12%	**22: 20.4%** **47: 10.2%** **61: 8.2%** **58: 8.2%**	**22: 16.1%** 32: 6.5% **33: 6.5%** **47: 6.5%**	**22: 13.2%** **47: 10.5%**	
C9orf72 (% of presence)		15.4%	6.2%	2.9%	7.1%	14%	0%	15.4%	5.3%	0%	0%	0%	4%	12.2%	3.2%	2.6%	**Genes**
SOD1		0%	0%	0%	3.6%	2.3%	2.4%	0%	0%	0%	0%	0%	0%	0%	6.5%	2.6%	
TARDBP		0%	3.1%	0%	0%	4.7%	0%	0%	5.3%	0%	0%	0%	0%	0%	0%	0%	
O3_Available (% of presence)		61.5%	62.5%	65.7%	57.1%	65.1%	63.4%	64.1%	63.2%	70.6%	55.6%	66.7%	76.0%	57.1%	61.3%	52.6%	**Environmental Characteristics**
O3_Ndays (Mean)		257.13 (0.27)	217.65 (0.46)	269.13 (0.29)	249.88 (0.33)	251.93 (0.33)	241.77 (0.38)	255 (0.36)	272.67 (0.26)	251.67 (0.41)	242.2 (0.28)	272 (0.31)	248.74 (0.3)	279.75 (0.21)	285.16 (0.2)	251.6 (0.35)	
O3_ratio_Medium		0.03 (0.84)	0.07 (1.98)	0.03 (0.86)	0.03 (1.29)	0.04 (1.33)	0.04 (1.05)	0.05 (1.64)	0.02 (1.18)	0.07 (1.69)	0.01 (1.4)	0.01 (1.1)	0.03 (1.16)	0.02 (1.11)	0.03 (0.94)	0.02 (1.19)	
O3_ratio_Poor		#VALUE!	0 (3.15)	0 (3.72)	0 (3.16)	0 (2.7)	0 (2.44)	0 (3.32)	0 (2.56)	0 (2.36)	0 (2.24)	#VALUE!	0 (3.38)	0 (4.14)	0 (2.61)	0 (4.47)	
PM10_Available (% of presence)		92.3%	75.0%	74.3%	100.0%	74.4%	80.5%	74.4%	94.7%	88.2%	88.9%	83.3%	84.0%	85.7%	77.4%	84.2%	
PM10_Ndays (Mean)		251.42 (0.26)	236.54 (0.45)	273.42 (0.3)	281 (0.24)	272.41 (0.24)	247.76 (0.38)	258.83 (0.34)	**288.11 (0.12)**	239.73 (0.35)	240.88 (0.26)	**289.4 (0.13)**	245.1 (0.34)	273.12 (0.28)	260.58 (0.3)	236.97 (0.43)	
PM10_ratio_Medium		0.06 (0.76)	0.07 (0.75)	0.06 (0.7)	0.06 (0.72)	0.07 (0.55)	0.07 (0.58)	0.06 (0.76)	0.07 (0.62)	0.05 (0.84)	0.08 (0.83)	0.04 (1.41)	0.06 (0.64)	0.08 (0.8)	0.07 (0.77)	0.07 (0.91)	
PM10_ratio_Poor		0.09 (1.07)	0.1 (0.9)	0.1 (0.94)	0.11 (0.89)	0.12 (0.71)	0.1 (0.88)	0.07 (1.15)	0.1 (0.89)	0.1 (1.18)	0.12 (0.96)	0.03 (1.25)	0.1 (0.97)	0.13 (0.72)	0.08 (0.96)	0.08 (1.06)	
PM25_Available (% of presence)		38.5%	62.5%	42.9%	50.0%	51.2%	46.3%	43.6%	52.6%	47.1%	55.6%	41.7%	52.0%	42.9%	32.3%	39.5%	
PM25_Ndays (Mean)		286.6 (0.03)	207.6 (0.53)	250 (0.28)	242.21 (0.33)	245.41 (0.35)	236.05 (0.37)	217.06 (0.42)	226.7 (0.45)	207.38 (0.5)	268.8 (0.09)	253 (0.27)	225.31 (0.36)	253.9 (0.25)	285.9 (0.04)	214.8 (0.5)	
PM25_ratio_Medium		0.05 (0.64)	0.06 (0.66)	0.07 (0.43)	0.07 (0.57)	0.07 (0.34)	0.07 (0.57)	0.06 (0.46)	0.07 (0.6)	0.06 (0.71)	0.04 (0.73)	0.03 (0.92)	0.08 (0.45)	0.08 (0.45)	0.07(0.53)	0.06 (0.51)	
PM25_ratio_Poor		0.11 (1.27)	0.22 (0.96)	0.22 (0.62)	0.19 (0.8)	0.28 (0.57)	0.17 (0.72)	0.15 (0.93)	0.14 (0.98)	0.21 (0.83)	0.22 (0.8)	0.02 (0.86)	0.21 (0.74)	0.24 (0.68)	0.22 (0.63)	0.19 (1.09)	
N02_Available (% of presence)		38.5%	12.5%	40.0%	39.3%	14.0%	24.4%	51.3%	42.1%	41.2%	33.3%	75.0%	32.0%	36.7%	51.6%	60.5%	

Mapping of Categories to Codes for: Affected Regions for Fine-Till-Nine Staging System & Patient disease history and occupation.

Fine_9_Affected_Groups (Initial or Final)	Code
Bulbar	0
Bulbar&Fine Motor	1
Bulbar&Fine Motor&Gross Motor	2
Bulbar&Fine Motor&Gross Motor&Respiratory	3
Bulbar&Fine Motor&Respiratory	4
Bulbar&Gross Motor	5
Bulbar&Gross Motor&Respiratory	6
Bulbar&Respiratory	7
Fine Motor	8
Fine Motor&Gross Motor	9
Fine Motor&Gross Motor&Respiratory	10
Fine Motor&Respiratory	11
Gross Motor	12
Gross Motor&Respiratory	13
No affected functional groups	14
Respiratory	15

Table A.4: Mapping for Fine-Till-9 Affected Groups

Disease History	Code
dyslipidemia	17
hypertension	22
hypertension;diabetes	27
hypertension;diabetes;dyslipidemia	32
hypertension;diabetes;dyslipidemia;cardiac_disease	33
hypertension;dyslipidemia	47
hypertension;dyslipidemia;cardiac_disease	50
hypertension;dyslipidemia;thyroid_disorder	56
hypertension;primary_neoplasm	58
hypertension;thyroid_disorder	61
primary_neoplasm	66
thyroid_disorder	69

Table A.5: Mapping for patient disease history

Occupation	Code
Administrative_and_commercial_managers	0
Building_and_related_trades_workers_excluding_electricians	3
Business_and_administration_associate_professionals	4
Business_and_administration_professionals	5
Cleaners_and_helpers	7
Drivers_and_mobile_plant_operators	10
Food_processing_wood_working_garment_and_other_craft_and_related_trades_workers	14
General_and_keyboard_clerks	15
Health_professionals	18
Labourers_in_mining_construction_manufacturing_and_transport	22
Legal_social_and_cultural_professionals	23
Market-oriented_skilled_agricultural_workers	25
Metal_machinery_and_related_trades_workers	27
Numerical_and_material_recording_clerks	29
Other_clerical_support_workers	30
Personal_service_workers	32
Sales_workers	36
Science_and_engineering_associate_professionals	37
Stationary_plant_and_machine_operators	39
Teaching_professionals	40

Table A.6: Mapping for patient occupation