This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

DYNAMITE: Integrating Archetypal Analysis and Process Mining for Interpretable Disease Progression Modelling

I. Trescato*, E. Tavazzi*, M. Vettoretti, R. Gatta, R. Vasta, A. Chiò, and B. Di Camillo, *Member, IEEE*

***Abstract*— DYNAMITE, an acronym for DYNamic Archetypal analysis for MIning disease TrajEctories, is a new methodology developed specifically to model disease progression by exploiting information available in longitudinal clinical datasets. First, archetypal analysis is applied to data organised in matrix form, with the aim of finding extreme and representative disease states (archetypes) linked to the original data through convex coefficients. Then, each original observation is associated with a single archetype based on their similarity; finally, an event log is created encoding the progression of disease states for each patient in terms of archetype states. In the last stage of the procedure, archetypal analysis is coupled with process mining, which allows the event log archetypes to be visualised graphically as sequences of disease states, allowing the clinical trajectories of patients to be extracted and examined. As a proof of concept, we applied the proposed method to data from a cohort of amyotrophic lateral sclerosis patients whose progression was monitored using the 12-item ALSFRS-R questionnaire. Without any a priori knowledge, DYNAMITE identified six archetypes clearly describing different types and severity of impairment and provided reliable clinical trajectories consistent with the prognosis of amyotrophic lateral sclerosis patients. DYNAMITE offers high interpretability at every stage of the analysis, which makes it particularly suitable for use in healthcare where explainability is paramount, and enables analysis of clinical trajectories at both individual and population levels.**

***Index Terms*— DYNAMITE, Archetypal Analysis, Clinical Data, Disease Progression, Disesase States, Longitudinal Dataset, Process Mining.**

I. Introduction

Archetypal Analysis (AA) is a powerful mathematical technique which can be applied to high-dimensional data space to identify a set of representative points (i.e., archetypes), which act as prototypical examples, encapsulating the essential characteristics of distinct subgroups or clusters within the dataset itself. Specifically, archetypes represent extreme behaviours that, when considered in combination, describe the entire data space.

This technique has found valuable applications in a variety of fields, including sociology (e.g., [1]), agronomy (e.g., [2]), economics (e.g., [3]), psychology (e.g., [4]), and healthcare (e.g., [5]).

In particular, healthcare stands out as a critical domain for leveraging this methodology to efficiently utilise clinical information towards the design of effective, personalised care: as healthcare systems increasingly transition towards digitalisation bringing a larger availability of electronic health records (EHRs) and other digital data sources, clinical research moves towards a systematic use of clinical data able, in principle, to accurately describe the status of heterogeneous populations of patients. Such clinical information often encompasses a wide range of variables, including demographic information, physiological measurements, medical history, and treatment outcomes. This complexity can pose a challenge when attempting to discern meaningful patterns and subgroups. AA addresses this challenge showing promising potential in unravelling complex patterns within clinical data by summarising the heterogeneity of patients, described through the collected variables, into a manageable number of archetypes, facilitating a deeper understanding of patients'/features' profiles and offering insights into underlying mechanisms of diseases.

In [6], the authors applied AA to gene expression data from small-cell lung cancer lines and tumours to identify phenotypic tasks consistent with cancer traits. In [7] and [8], AA was used to examine the distribution of molecular phenotypes, using as input the classification scores obtained from an ensemble learning model on a set of histological and RNAseq samples, respectively. In [9], the authors demonstrated that, when applied to genomic data, AA can be used to estimate genetic clusters that have a similar structure to those obtained with standard Bayesian or likelihood-based population clustering methods, with clear advantages in terms of computational efficiency. In [10], AA was employed to provide a compact representation of the current catalogue of single-base substitution signatures extracted from 2,780 genomes. AA also has numerous applications in the identification of visual field loss (VF) patterns starting from VF testing data of patients with glaucoma [11], [12], optic neuritis [13], [14] and idiopathic intracranial hypertension [15]. In [16] AA were applied to anthropometric data to show the efficacy of this methodology in automatically discriminating between genders, starting from skeleton measurements only. In [17], the authors used AA to characterise a cohort of subjects affected by a neurodegenerative disease, starting from their clinical data collected at diagnosis.

While all these applications make use of static information only, such as biopsies, omics, or clinical tests, healthcare data often includes a more detailed description of a patient's medical condition than just a static point, frequently consisting of multiple observations spanning from an initial assessment, such as study entry or hospital admission, to a specific clinical event (e.g., survival or hospital discharge). In this case, the available information can be even more complex, consisting not only of data that are heterogeneous from a feature type (i.e., continuous, categorical, or ordinal) but also from a temporal point of view. The data acquisitions can indeed vary from subject to subject according, for instance, to the patient monitoring

Submission date: 17th July, 2024. This research has been partially supported by the University of Padova project C94I19001730001 "Deconstruct and rebuild phenotypes: a multimodal approach toward personalised medicine in ALS (DECIPHER-ALS)", by the Italian Ministry of Health grant (Ricerca Finalizzata) RF-2016-02362405 "Identification of genetic and environmental determinants of onset and progression of ALS (INITIALS)", by the Italian Ministry of Education, University and Research grant for Research Projects of National Relevance (PRIN) 2017SNW5MB, by the initiative "Departments of Excellence" of the Italian Ministry of Education, University and Research (Law 232/2016), and by the Horizon 2020 project BRAINTEASER (Bringing Artificial Intelligence home for a better care of amyotrophic lateral sclerosis and multiple sclerosis) funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No. GA101017598.

*These authors contributed equally.
I. T., E. T., M. V., and B. D. C. are with the Department of Information Engineering of the University of Padova, Padova, 35131, Italy (emails: isotta.trescato@phd.unipd.it, erica.tavazzi@unipd.it, martina.vettoretti@unipd.it, barbara.dicamillo@unipd.it); R. G. is with the Department of Clinical and Experimental Sciences of the University of Brescia, Brescia, 25121, Italy (email: roberto.gatta@unibs.it); R. V. and A. C. are with the Neuroscience Department "Rita Levi Montalcini" of the University of Torino, Torino, 10126, Italy (emails: rosario.vasta@unito.it, adriano.chio@unito.it); B. D. C. is with the Department of Comparative Biomedicine and Food Science of the University of Padova, Legnaro (PD), 35020, Italy (email: barbara.dicamillo@unipd.it).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

schedule or the specific variable granularity – and be asynchronous with respect to the population. This results in a sampling grid that, in general, may vary both along the patient-specific observation period and from patient to patient [18]. When analysing dynamic contexts, all these data characteristics must be taken into account.

By making use of methodologies that can adequately exploit longitudinal clinical information, the study of clinical conditions can be extended to characterise, for example, patients in terms of progression trajectories and similarity of prognosis. When modelling disease trajectories, it has to be taken into account that, in general, each individual follows a distinct path, influenced by a myriad of factors including genetics, lifestyle, comorbidities, and treatment regimens. Therefore, deciphering the underlying patterns within this diversity requires advanced analytical techniques capable of accommodating non-linearity, variability, and irregularity in the data.

In this paper, we propose DYNAMITE (DYNamic Archetypal analysis for MIning disease TrajEctories), a new method based on an innovative use of AA for unravelling disease progression patterns and outlining patient trajectories starting from a longitudinal clinical dataset.

Acting on a series of consecutive clinical observations (visits), instead of modelling patient profiles, we aim at encapsulating in the archetypes the extreme disease states within which patients may move as the course of their illness evolves. The combination of these archetypes characterises the range of potential dynamic states observed within the study population. As a result, for each subject, the sequence of their visits can be mapped on a sequence of archetypes, which define the evolution of disease progression through extreme representative states. We then adopt Process Mining (PM) to represent the sequence of disease states transitioned by the patients and to inspect the distinct patterns of progression over the whole patient population.

Originating in the context of business process management and then spreading to other contexts – including, recently, healthcare [19], [20] - PM is a relatively young analytical discipline that provides methodologies to represent and study processes [21]. PM techniques require as input an Event Log (EL), that is, a sequence of ordered events (or activities) referred to a set of cases, each labelled with their occurrence times and possibly enriched with a set of optional attributes that characterise the case or the activity. Specifically, the sequence of events for each case is called the trace in PM. Here, the traces are the sequences of archetypes mapped for each patient. PM provides different techniques and tools, mainly aimed at (i) mining the processes that produced the input data (process discovery) [22], (ii) assessing the conformity of an EL with regards to a given process, or vice versa (conformance checking) [23], and (iii) improving the efficiency of a process via diagnostic and restorative strategies (process enhancement) [24]. In the context of dynamic clinical data, PM can therefore be particularly valuable for tracking patient trajectories over time in terms, for instance, of sequence of interventions, transfers between wards, or - as in our case - transitions between disease states.

To show a proof of concept of DYNAMITE, we present the analysis of a longitudinal clinical dataset extracted from a register of amyotrophic lateral sclerosis (ALS) patients. Our methodology allows the identification of six archetypes that describe the extreme health states experienced by the patients as the disease progresses, clinically corresponding to the stand-alone or the combination of functional impairments caused by ALS. By analysing the mined sequence of archetypes, we were able to identify the most frequent patterns of progression in terms of transitions and percentages of subjects experiencing the sequences of clinical states, up to the occurrence of the survival outcome.

Our approach proves useful in practice to model the history of patients with heterogeneous progressions in an effective and communicative manner, managing the complexity of the clinical data and ensuring good interpretability.

II. Method

With the aim of modelling the evolution of the clinical condition in a study population starting from a collection of longitudinal data, we propose a four-step procedure:

A) organisation of the patients' dynamically-collected data as a longitudinal clinical dataset in a matrix form;
B) use of archetypal analysis (AA) on the longitudinal dataset to identify the most extreme behaviours of the patient visits and characterisation of the resulting archetypes;
C) definition of the clinical progression trajectories, by associating each patient's observation with the most representative archetype, representing their punctual, subsequent *disease state*;
D) employment of process mining (PM) techniques to represent the clinical progression trajectories consisting of the sequence of disease states.

In the following, we characterise these analytics steps.

A. Organisation of a longitudinal clinical dataset in a matrix form

Longitudinal clinical datasets are collections of clinical facts referring to a specific cohort of individuals monitored consecutively at multiple points in time.

Mathematically formalising, given a cohort of N samples s_k with $k = 1, ..., N$ corresponding to the observed subjects, each subject contributes with m dynamic features collected over consecutive time points t_{k,j_k}, with j_k possibly different in sampling time and number for each subject k. Figure 1 shows (a) the tri-dimensional structure of the data upon collection, and (b) the corresponding $n \times m$ matrix X where the observations have been stacked consecutively.

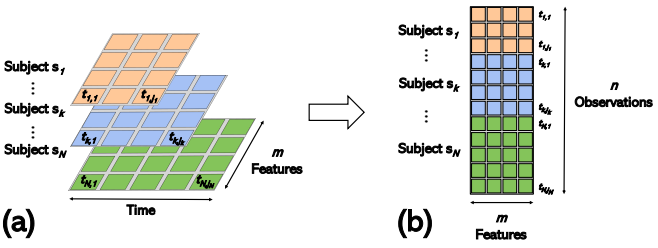

Fig. 1. Structure of longitudinally collected clinical data: (a) three-dimensional representation with dimensions corresponding to patients, time, and variables; (b) matrix form.

B. Computation of Archetypal Analysis and inspection of the archetypes

1) Archetypes computation: As a modelling technique, we employ the AA on the longitudinal dataset X collecting the clinical information of the study population, where the m columns correspond to the set of variables dynamically collected and the n rows to the patients' observations (or visits).

In detail, given an $n \times m$ matrix X representing a multivariate dataset, where n is the number of observations and m is the number of variables, and provided a number k of archetypes, AA allows determining a $k \times m$ matrix Z of archetypes such that:

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

1) the data are best approximated by convex combinations of the archetypes Z, *i.e.*, they minimise the residual sum of squares (RSS):

$$RSS = ||X - \alpha Z^T||_2, \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1 \tag{1}$$

2) the archetypes are convex combinations of the data points:

$$Z = X^T \beta, \quad \beta_i \geq 0, \quad \sum_i \beta_i = 1 \tag{2}$$

where $||\cdot||_2$ represents the matrix norm, α are the coefficients of the archetypes and β are those of the dataset, respectively [25].

The AA algorithm starts from a set of k randomly chosen points in the features' space, and iterates between finding the best α for given archetypes Z and finding the best archetypes Z for given α.

The learning process stops either when the residual sum of squares (RSS) reduction is smaller than a defined threshold or when the maximum number of iterations set is reached. The AA training procedure, therefore, will always converge, but to minimise the possibility of converging to a local minimum, instead of a global one, it is recommended to restart the training multiple times.

Another aspect to keep in mind is that, as in other unsupervised analysis techniques, there is no method for determining in advance the optimal number of k archetypes. One possible, heuristic solution is to test several k values in a defined range and choose the optimal value by detecting the elbow (that is, a sharp change of direction) in the scree plot reporting the k values on the x-axis and the RSS values on the y-axis [25].

Archetypes, by construction, lie on the boundary of the convex hull of the data and thus can be easily influenced by outliers. To mitigate the effect of possible outliers and avoid incorrect or skewed results, we employed the `archetypes` R package [16], which implements the weighted and robust archetypal analysis proposed by Eugster et al. [26]. In this method, the original algorithm is adapted to reduce the influence of outliers by using M-estimators instead of least squares estimators when performing the optimisation procedure. Specifically, instead of minimising the Euclidean norm of the residuals $R = (X - \alpha Z^T)$, i.e., $\min||R||_2$ where large residuals have large effects (see Eq. 1), the M-estimators try to reduce the effect of outliers by replacing the squared residuals by another function $\rho(\cdot)$ less increasing than the square.

2) Archetypes representation: Once the archetypes have been computed, Principal Component Analysis (PCA) is used as the data visualisation technique to graphically check their position in the data space.

In such representation, archetypes are expected to be on (an approximation of) the convex hull of the data, according to their definition. To visualise how each archetype is characterised in terms of variables, radar charts are then employed. Specifically, each variable is reported on an axis where: for continuous variables, their position on the axis corresponds to their value in the min-max range, computed on the whole dataset, while for categorical and ordinal variables distinct values on the axis correspond to the feature levels.

C. Definition of the clinical progression trajectories

1) Association of each observation with the most representative archetype: After performing AA, each observation x_j corresponds to a linear combination of the k archetypes z_i, as reported in Eq. 3:

$$\begin{gathered} x_j = \alpha_{j,1} z_1 + \cdots + \alpha_{j,k} z_k \quad i = 1, \ldots, k \\ \alpha_{j,i} \geq 0, \quad \sum_i \alpha_{j,i} = 1 \end{gathered} \tag{3}$$

where a higher coefficient $\alpha_{j,i}$ means a higher resemblance between the observation x_j, i.e., visit, and the archetype z_i.

With the aim of describing the progressing status of the patients as a sequence of prototypical conditions, we can associate each observation with the archetype that best represents it, according to one of the following criteria [27]:

- *crisp rules*: each observation is associated with the nearest archetype, that is, the archetype z_w whose coefficient $\alpha_{j,w}$ is the biggest. Formally:

$$x_j \coloneqq z_w \mid w = \underset{i}{\operatorname{argmax}}\, \alpha_{j,i} \tag{4}$$

- *fuzzy rules*: each observation is associated with the archetype(s) z_w whose coefficient $\alpha_{j,w}$ are bigger than a given threshold τ set by the user. Note that a threshold $\tau > 0.5$ is needed for a univocal association; however, with such a threshold, some visits may not turn out to have a valid association. Formally:

$$x_j \coloneqq z_w \mid \alpha_{j,w} > \tau \tag{5}$$

In our method, we adopt the crisp assignment and represent each visit with the archetype z_w with the higher α_j, w coefficient, i.e., the most similar one.

After the assignment, for each subject, we obtain an ordered sequence of matching archetypes corresponding to the evolution of the disease states.

2) Sequence reduction to remove multiple consecutive events: According to the archetype assignment, consecutive visits might be associated with the same archetype. With the aim of focusing on the progression from one disease status to another, when the same archetype is assigned to several consecutive visits we reduce the sequence of assigned archetypes by keeping only the first observation with the corresponding sampling time. This maintains the sequence of disease states, avoiding repetitions that would correspond to confirmation that, over time, the patient is still associated with that specific extreme behaviour, a piece of information that – for the sake of capturing the timing of the transition from one disease state to another – does not add knowledge to the data.

In Figure 2, we report a practical example of the procedure described above for obtaining the reducing sequence of disease status starting from the longitudinal observations of a generic subject s_k.

D. Characterisation of the clinical trajectories through Process Mining

As a next and final step, we use PM to characterise the clinical trajectories obtained in terms of the succession of disease states.

Here, we employ a process discovery algorithm, namely the Care-Flow Miner (CFM) [28], to represent and characterise the patients' clinical trajectories. The CFM algorithm was already used in the literature to model healthcare processes in different clinical domains, such as diabetology [29], oncology [28], [30], and neurology [31]. With respect to the PM nomenclature introduced above, here the traces correspond to the reduced sequence of disease states obtained with the procedure detailed so far, and the EL is the combination of all traces together. For each case (here subject), the events correspond to the archetypes associated with each visit, while the timestamps are the times of each considered observation. All traces together, that is, the whole longitudinal dataset reduced to its corresponding archetypal representation, constitute the EL.

Starting from an EL, the CFM algorithm returns the process that generated the data in a graphical form in which events are represented as nodes and transitions between events are reported as directed arcs. Starting from a *root* node, each EL trace contributes to building a branch of a tree, with the top level (i.e., the first one after the root) representing the first event of each trace and the subsequent levels corresponding to the next and ordered events of

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

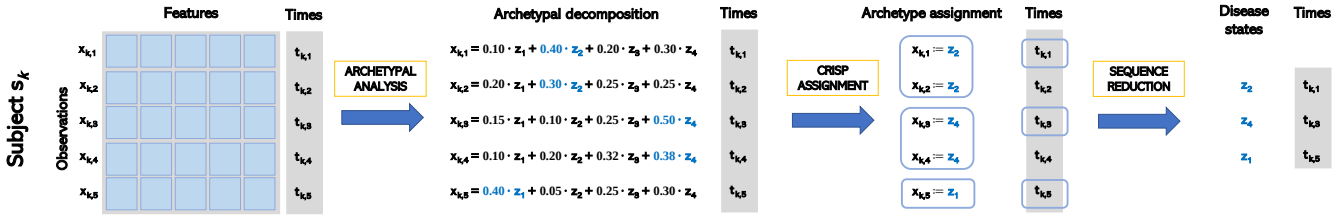

Fig. 2. Method workflow.

each trace. The process represented carries a label on each node with the name of the associated event and some details such as the number of cases (here, subjects) who passed through it or statistics on the time taken to reach it from the first event in the trace. Through this representation, clinical trajectories can be inspected in terms of the cardinality of subjects following different paths, probabilities or transition times between one event and the next. When applied on different datasets, this methodology allows comparison of trials constructed on different datasets, highlighting differences and similarities in terms of trajectories in different cohorts.

For this part of the analyses, we employed the `pMineR` R package [32], a library that offers implementations of the traditional PM algorithms specifically enriched to address the needs of the healthcare domain, integrating them with survival analysis and features to explore differences among clinical pathways. This implementation of the CFM also allows tuning some parameters of the algorithm, such as a threshold on the minimum number of instances considered significant by the user to visualise a transition, which allows limiting the so-called *spaghetti effect* typical of this kind of representation. This effect, determined by the possible high variability of the trajectories represented in the data, corresponds to a potentially large number of represented infrequent paths that increase the complexity of the process without being as informative.

III. Case Study

A. Longitudinal clinical dataset

As a case study for the proposed methodology, we selected a clinical register of Amyotrophic Lateral Sclerosis (ALS) patients, namely the Piemonte and Valle d'Aosta Register for ALS (PARALS) [33]. ALS is a rare neurodegenerative disease that affects motor neurons, reducing the ability to control voluntary movements. Its manifestations are highly heterogeneous in both symptoms and progression patterns, typically leading to death within 3-5 years from the onset, mostly due to breathing muscles paralysis. A wide-world recognised tool to measure disease progression is the Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) [34], especially in the revised version (ALSFRS-R) [35]. ALSFRS-R scale is a questionnaire composed of 12 questions, regarding 5 different domains:

- questions 1-3 are linked to bulbar (BU) ~~domain~~ impairments;
- questions 4-5 are linked to upper limbs (UL) impairments;
- questions 6-7 are linked to trunk (TR) ~~domain~~ impairments;
- questions 8-9 are linked to lower limbs (LL) ~~domain~~ impairments;
- questions 10-12 are linked to breathing (BR) impairments.

Each item is evaluated by assigning a score between 0 and 4, where the highest value is associated with "normal" and a lower value is attributed as the impairment progresses. ALSFRS-R total score, indeed, ranges between 0 and 48 with 48 indicating a healthy person.

To characterise and study the progression of ALS, it is crucial to consider the trajectories of the disease. Patients with ALS commonly experience a worsening of their condition over time, with an increasing number of domains being affected. This progression can be observed through changes in the ALSFRS-R scores over time. By analysing longitudinal data from multiple visits, it becomes possible to track the evolution of the disease by examining how ALSFRS-R scores are modified.

In this context, therefore, the first aim is to use DYNAMITE to identify meaningful archetypes representing different disease states and associate each patient visit with the corresponding archetype. By studying the sequence of archetypes across visits for the considered subjects, it becomes then possible to map and analyse the trajectories of disease progression in ALS patients.

The employed dataset consists of 8392 visits referred to 923 ALS patients extracted from the PARALS register. The extraction, carried out in November 2021 within the framework of the project 'Deconstruct and rebuild phenotypes: a multimodal approach towards personalised medicine in ALS' (Research Projects of National Significance, Italian PRIN call 2017), includes patients diagnosed between 2007 and 2015 and consecutively enrolled in the registry. No inclusion/exclusion criteria were adopted. The diagnosis of ALS was assessed according to the revised El Escorial diagnostic criteria [36] by expert neurologists. As the data we used came from a real clinical registry rather than a clinical study, a rather long follow-up was generally recorded for the subjects. In the study population, the number of visits ranged from 1 to 44, with a median of 8 (IQR 4-12). The median interval between visits is 75 days (IQR 52-105), and the median clinical follow-up, calculated here as the time interval between the first and last recorded ALSFRS-R assessment, is 878 days (IQR 346-1254). For each visit, the scores of the 12 items of the ALSFRS-R scale are recorded, together with the corresponding collection date. When available, the date of tracheostomy or death was used to code the survival event for each subject.

The PARALS register database is anonymised and treated according to the Italian Data Protection Code. Written informed consent to participate in the study was obtained from all the patients or their legal representatives. This study was approved by the ethical committees Azienda Ospedaliera Universitaria City of Health and Science of Turin (number 004462, June 10, 2010). A waiver was obtained for patients included in the register before 2010. The Piedmont regional government also recognised PARALS as a 'Registry of High Sanitary Interest' (Regional Law, April 11, 2012, number 4). Accordingly, the PARALS can access databases owned by the regional administration to obtain clinical information about ALS patients from public and private hospitals, and general practitioners. All study protocols and procedures were conducted in accordance with the Declaration of Helsinki.

B. Computation of Archetypal Analysis and inspection of the archetypes

For the archetypes computation, only the 12 columns representing the ALSFRS-R questions were included, considering each visit as an independent observation. To avoid the archetypes being influenced

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

by outliers, which are frequent in clinical datasets, robust archetypes were trained ("*robustArchetypes*" function). The training procedure was repeated 15 times (*nrep =15*), as it was found to be a fair trade-off between the soundness of the procedure and computational time. The stopping criteria were set equal to the default values: RSS reduction smaller than $1.490e-08$ (corresponding to the square root of the machine epsilon for double-precision floating-point numbers) and a maximum number of iterations equal to 100 [16]. For each iteration, all integer numbers of archetypes k between 2 and 10 were tested and the optimal number of archetypes was chosen by detecting the elbow on the scree plot reporting on the x-axis the value k and on the y-axis the RSS. To reinforce the choice of the best k, the minimum, mean, and standard deviation of the RSS over the 15 repetitions, for each k, were also analysed [27].

Figure 3 shows the RSS of the best model for each step on the y-axis, and the number of archetypes k on the x-axis.

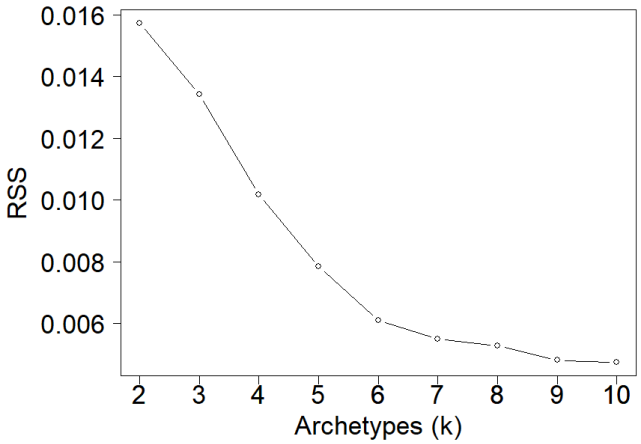

Fig. 3. Scree plot obtained in the ALS case study, reporting the number of archetypes k on the x-axis and the RSS values on the y-axis.

Increasing k up to 6 seems convenient in order to achieve an improvement in RSS as displayed in Figure 3.

The mean, minimum, and standard deviation of the RSS over all repetitions were also evaluated to identify the k value that allows a good trade-off between complexity and performance. Table I reports the RSS for each repetition (in the columns) and for each k (by rows). This analysis confirmed k = 6 as the optimal value of archetypes for the case-study dataset.

The α coefficients, which characterise the 6 archetypes in the 12 ALSFRS-R questions were derived for the repetition that had the lowest RSS. Figure 4 represents the six derived archetypes in six radar plots, useful for observing the differences between them. In all these plots, each axis corresponds to an ALSFRS-R item: the inner level of the plot represents a value of 0 for that score, indicating an impaired state, while the outer circle corresponds to the maximum score of 4, representing a fully functional state. The blue shape on the graph characterises the archetypes, showing for each of the 12 ALSFRS-R items the value of the corresponding position on the axis.

Notably, even though AA is an unsupervised method, the analysis resulted in finding archetypes that trace the domains of the ALSFRS-R. Indeed, the identified archetypes show consistency with the five domains investigated with the score, so that the scores for the different items related to each domain concordantly indicate an impaired or a preserved domain. For clarity, each archetype was renamed according to the eventually impaired domains. A domain, as defined in section III-A, was considered impaired if and only if the mean of the scores of the question that constitute the domain itself was ≤ 2. The archetype labelled as "none" is characterised by high scores in all 12 questions, thus representing subjects who, although diagnosed with ALS, have no significant impairments. Conversely, the archetype "all" has low scores in every question and thus represents subjects who are highly impaired in every assessed domain. The other identified archetypes represent disease states between these two. Specifically:

- The archetype "BU" presents a well-conserved breathing domain and an evident impairment on the bulbar domain. It also displays moderate scores in items 4 to 9, related to limb impairments.
- The archetype "BR" is characterised as having high scores in all the questions, except for the breathing domain.
- The archetype "LL" displays high scores for items 1 to 5 and 10 to 12, a moderate impairment for the trunk domain, and a relevant impairment on lower limbs.
- The archetype "UL_TR_LL" has the bulbar and the breathing domains well preserved, but the low scores of questions 4 to 9 highlight impairments in the upper limbs, trunk, and lower limbs domains.

We also characterised each archetype in the original space of features by computing their ALSFRS-R total score (see Table II). Interestingly, different archetypes correspond to the same or similar ALSFRS-R total scores: the fact that different archetypes, characterised by different (sets of) impairments, map onto the same total scores suggests that this representation is more effective in distinguishing between different disease states than the direct use of the total score.

Another descriptive representation of the archetypes is obtained by applying the PCA and it is reported in Figure 5. PCA allows obtaining a 2D representation of the original 12-dimensional space. In the plot, it is possible to observe the distribution of the visits (light blue dots) and of the archetypes (blue diamonds). It should be noted that the archetype "none", on the left, is opposite to the archetype "all", which is on the right, as expected. Similarly, the archetype with the bulbar domain impaired ("BU") is on the top part of the graph, while the one associated with impairment in the lower limbs ("LL") is on the lower part. This configuration is consistent with what is known about ALS in the literature, where bulbar and spinal impairments are highly distinctive and characteristic conditions right from the onset [37].

C. Definition of clinical progression trajectories

1) Association of each observation with the most representative archetype: As described in section II-C, the association of each visit with the most representative archetype was based on the α coefficients and followed the nearest prototype assignment rule. More in detail, the procedure consists of 4 steps as follows:

1) The α coefficients, that describe each visit in the archetypes' space, were derived: doing so, a 8392×12 dataset was obtained.
2) For each row, the highest coefficient was selected as a mean to identify the nearest archetype (crisp rules).
3) A new dataset was created, with a row for each visit and three columns. As required to build an EL, the first column reported the patients' id, the second column reported the visit date, and the last column reported the name of the nearest archetype (one among "none", "all", "BU", "LL", "BR", and "UL_TR_LL").
4) Lastly, for each subject, a row was added to specify whether he or she was dead. If, for a given patient, a death date was reported, the second column was filled with the death date and the third column with the string "Dead". Otherwise, if a subject was not reported to be dead, the censoring time was set to the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

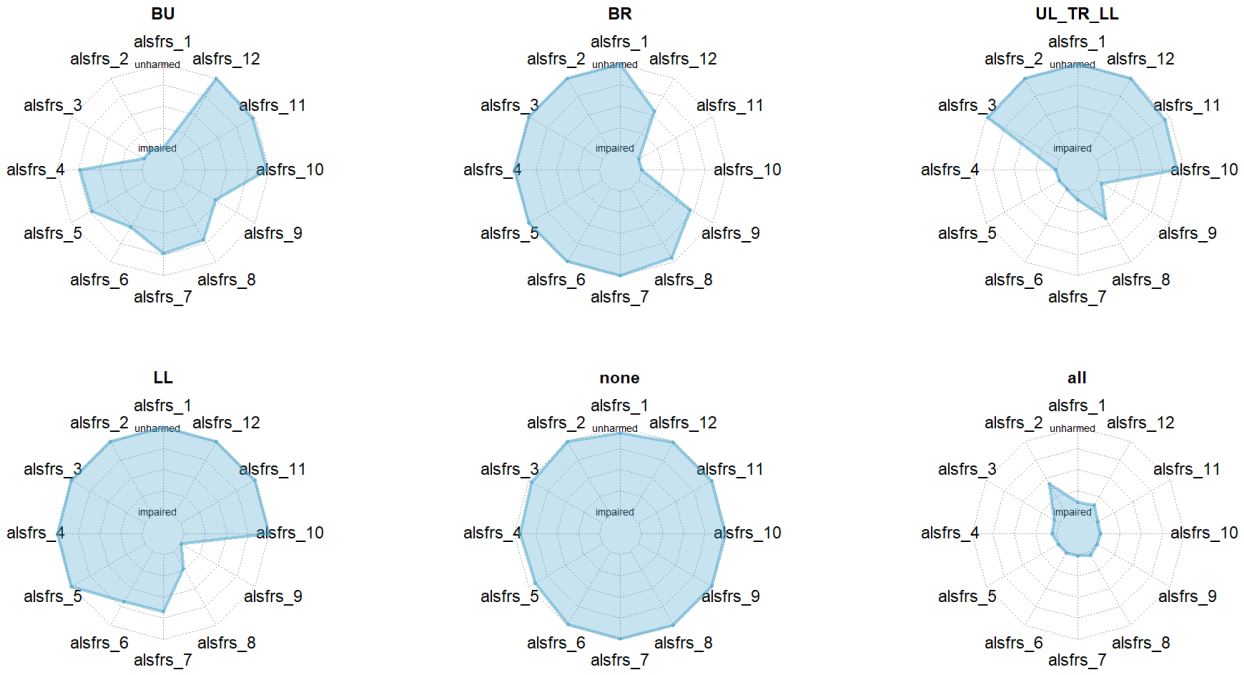

Fig. 4. Radar plot of the archetypes obtained in the ALS case study.

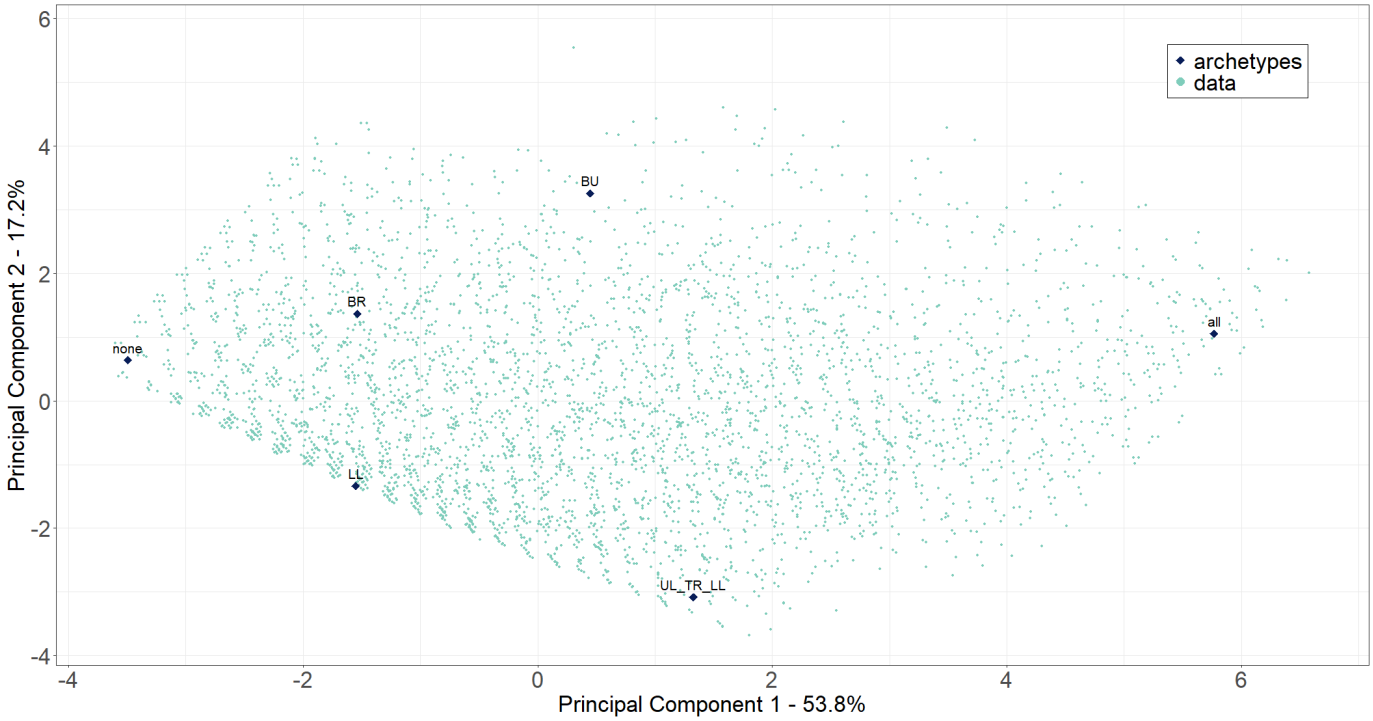

Fig. 5. PCA of ALS data and derived archetypes.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

	Rep1	Rep2	Rep3	Rep4	Rep5	Rep6	Rep7	Rep8	Rep9	Rep10	Rep11	Rep12	Rep13	Rep14	Rep15	Mean	Min	SD
k=2	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0
k=3	0.014	0.014	0.013	0.014	0.014	0.013	0.014	0.013	0.014	0.014	0.013	0.014	0.014	0.013	0.014	0.014	0.013	0
k=4	0.013	0.010	0.010	0.010	0.010	0.010	0.010	0.01	0.010	0.010	0.010	0.014	0.013	0.010	0.013	0.011	0.010	0.001
k=5	0.011	0.010	0.008	0.013	0.012	0.010	0.010	0.008	0.009	0.010	0.009	0.009	0.009	0.009	0.009	0.010	0.008	0.001
k=6	0.008	0.010	0.006	0.010	0.006	0.006	0.008	0.009	0.008	0.007	0.009	0.010	0.010	0.008	0.010	**0.009**	**0.006**	**0.002**
k=7	0.010	0.006	0.008	0.010	0.006	0.006	0.012	0.010	0.006	0.006	0.005	0.009	0.006	0.007	0.006	0.007	0.005	0.002
k=8	0.006	0.010	0.007	0.010	0.009	0.008	0.008	0.010	0.005	0.007	0.006	0.008	0.008	0.010	0.012	0.008	0.005	0.002
k=9	0.006	0.011	0.006	0.005	0.005	0.007	0.011	0.007	0.005	0.009	0.006	0.006	0.010	0.006	0.011	0.007	0.005	0.002
k=10	0.005	0.005	0.005	0.013	0.005	0.010	0.005	0.015	0.005	0.005	0.005	0.011	0.012	0.005	0.006	0.007	0.005	0.004

TABLE I

VALUES OF THE RSS OBTAINED IN THE ALS CASE STUDY FOR EACH REPETITION (IN THE COLUMNS) AND EACH NUMBER OF ARCHETYPES k (IN THE ROWS), ALONG WITH THEIR MEAN, MINIMUM, AND STANDARD DEVIATION FOR EACH k. IN BOLD, THE SELECTED k.

Identified archetype	Total ALSFRS-R score
none	47
LL	38
BR	38
BU	28
UL_TR_LL	26
all	4

TABLE II

ARCHETYPES IDENTIFIED IN THE ALS CASE STUDY AND THEIR CORRESPONDING TOTAL ALSFRS-R SCORE.

Event (archetype or survival)	N occurrences	% on the total
none	2234	26.62
all	1216	14.49
BU	982	11.70
BR	513	6.12
LL	1635	19.48
UL_TR_LL	1812	21.59
Dead	732	79.31
Censored	191	20.69

TABLE III

OCCURRENCES OF EVENTS IN THE FINAL EL OF THE ALS CASE STUDY: THE TOP SECTION REPORTS THE VISIT FREQUENCIES AND PERCENTAGES FOR THE SIX ARCHETYPES, WHEREAS THE BOTTOM SECTION OUTLINES THE FREQUENCIES AND PERCENTAGES OF DEATH AND CENSORING IN RELATION TO THE TOTAL SUBJECT COUNT.

day after the last recorded visit and reported in the second column, and the third column was filled with "Censored".

Note that the visit dates and the death/censoring dates were not used to derive the archetypes, but were available in the original datasets and are useful for the following PM analysis.

2) Sequence reduction to remove multiple consecutive events: To avoid redundancy and focus on transitions between different disease states, sequence reduction was applied to the event log as described in Section II-C.2. The top section of Table III shows the frequency and percentages of visits for the six archetypes, while the bottom section displays the frequency of censoring and death along with their percentages relative to the total number of subjects.

D. Characterisation of the clinical trajectories through Process Mining

Figure 6 represents the graph obtained with the CFM algorithm on the EL considered. All traces start from the conventional *root* node and are represented as descending collections of nodes connected by arcs. On each arc, the percentage of subjects transitioning between the two nodes is reported, computed based on the total number of subjects in the node and the total number of subjects in the dataset.

The first step, i.e., the first disease state that a patient can be attributed to, appears to be any archetype except the one with all domains impaired. From the graph, it is quite evident that the shorter traces belong to those subjects who, in the first available assessment, present a bulbar or breathing impairment (archetypes "BR" or "BU", respectively), or those with three domains impaired (archetype "UL_TR_LL"). A shorter trace can be interpreted as the occurrence of fewer distinct intermediate disease states before the survival event.

Subjects who, at the first assessment, have a lower limb impairment (archetype "LL") display differentiated prognosis steps, with the majority moving to the archetype with impairment in the lower limbs, trunk and upper limbs (archetype "UL_TR_LL"). Lastly, most of the subjects in our data at the first assessment had no relevant impairments and therefore were assigned to the "none" archetype as the first disease state. Then this group shows a highly differentiated prognosis, with a minor percentage of subjects having death, censoring, or breathing impairment as the subsequent status. The remaining subjects are equally distributed among those who have only bulbar impairment, only lower limb impairment, or combinations of lower limb, trunk, and upper limb impairments.

While analysing the graph, it is crucial to keep in mind that thresholds have been set to avoid a spaghetti plot, which prevents the representation of less frequent traces. Specifically, we acted on the setting of the two parameters *max depth* and *minimum number of subjects* of the CFM visualisation function, set equal to 6 and 10, respectively. A clear sign of the use of such thresholds is that the sum of the percentages displayed does not reach 100%.

IV. DISCUSSION AND CONCLUSION

In this work, we presented DYNAMITE, an innovative methodology that combines AA and PM to identify representative disease states and mine disease progression trajectories from longitudinal clinical data. The proposed method is particularly suitable for clinical applications in which the same features, such as questionnaires or clinical tests, are assessed over multiple visits to monitor the progression of the disease. As a proof-of-concept, we demonstrated the applicability and usefulness of DYNAMITE on a longitudinal clinical dataset of ALS patients. The method is based on a four-step procedure, which involves the data preprocessing needed to organise the dynamically collected data in matrix form (Section II-A), the employment of AA to identify extreme disease states and to characterise them (Section II-B), the creation of an EL through the association of each observation of the patient with the closest archetype (Section II-C), and lastly the analysis of the patients' trajectories among the archetypes with PM techniques (Section II-D).

It should be noted that the proposed methodology guarantees straightforward interpretability at every step. The identified archetypes can indeed be characterised in the original space of features in three different ways:

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

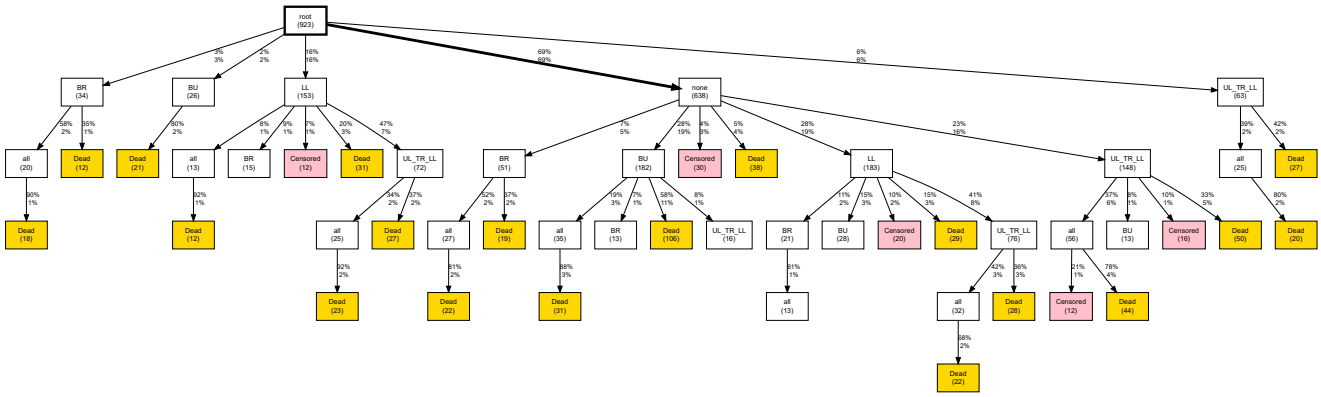

Fig. 6. CFM graph of the ALS case study: each node reports the total count of patients passing through it (round brackets), while the edges show the percentage of patients that continue to the next node in relation to both the preceding node (top) and the overall population (bottom). Survival events are highlighted with yellow (dead) and pink (censored). The graph is filtered to display only those pathways that have been followed by at least 10 subjects, with a maximum depth of 6.

- through coefficients, which represent them within the original dataset as additional states constituting the extreme states of disease;
- through radar plots, to highlight and capture their peculiarities and to facilitate comparisons among them;
- through PCA, enabling the representation of both the original data points and the archetypes in a reduced space (e.g., 2D or 3D).

To ensure that the computation of archetypes was not excessively impacted by possible clinical outliers, we employed a robust implementation of AA that provides unbiased results based on the majority of data points [26].

By combining the AA with a graphical technique like PM, DYNAMITE makes it possible to effectively analyse the sequences of archetypes associated with the consecutive visits of a subject or population, after organising them in the form of an EL. The inspection of the obtained traces through the CFM not only allows the analysis of the progression trajectory of a single subject but also highlights the prevalent trajectories within the population.

Regarding the arrangement of archetype sequences as traces, it should be noted that by default, our approach condenses the EL data to eliminate successive identical events. Although effective in terms of mining and visualisation, this aggregation step results in a loss of part of the temporal information in the data. Specifically, one loses the confirmation that, in the case of successive visits attributed to the same archetype, the subject was traced as remaining in the same disease state consecutively. It must be acknowledged, however, that quantifying the permanence of the same disease status is not straightforward. Although the methodology chosen for the representation of trajectories, i.e., PM, provides tools for analysing, at the individual or population level, the transition times between one state and the next, the limitation of this analysis lies in the very nature of the data. The evolution of a clinical condition is, in fact, a continuum of states, sampled only at the time of visits. Despite the continuous nature of the condition, this sampling takes place on a discrete and non-uniform grid. This implies that it is not possible to reconstruct the exact moment of the switch from one state to the next from the data, being limited to the clinician's recording of a different symptomatology or characterisation of the subject. However, in other contexts where sampling may be more regular or dense, it may be worthwhile to choose whether or not to aggregate identical and consecutive archetypes. For this reason, we have released an implementation of DYNAMITE that allows the user to choose whether or not to aggregate identical and consecutive archetypes through an option that removes or retains autoloops in the CFM graph.

A further observation on the temporal aspects of the original real-world data and their effect on trajectories concerns the actual availability of dynamic data, for which borderline situations may exist. In particular, there may be subjects for whom only one visit is available, or subjects with several visits but all associated with the same archetype. This situation also occurred in our case study, where subjects with only one ALSFRS-R assessment or with all visits mapped to the same archetype gave rise to traces composed of only two events, of the type: archetype associated with visit(s) - death/censure event. While it is true that traces of this type do not provide information on the progression of the subject in the strict sense, they do contribute to characterising the population under study in the initial phase of the trajectories. Furthermore, since all traces include a survival event, they allow survival analyses to be carried out on the entire sample or on sub-cohorts with specific characteristics (e.g. based on the static characteristics of the subjects, or on the first archetype(s) characterising progression).

As a comparison with other possible implementation choices, we evaluated the use of AA in our methodology against other techniques capable of summarising representative behaviour of the data. Specifically, we replaced in our method what was the step of calculating the archetypes and assigning each visit to one of them with clustering k-means with multiple random restarts and a number of clusters equal to the number of identified archetypes, i.e., 6. We then used centroids to represent and characterise disease states, instead of archetypes, and modelled patients' trajectories through them. Figure 7 displays the radar plots of the centroids obtained: compared with the archetypes reported in Figure 4, we can observe that the shapes of the centroids resemble those of the archetypes but appear more moderate, i.e., less prone to assume values tending towards the minimum or maximum of the range of ALSFRS-R items. That is, the disease states represented by the archetypes are more extreme than those defined by the centroids, which instead reflect more average conditions.

This was somewhat expected, since, by construction, the archetypes lie on the convex hull of the data while the centroids are positioned in the centre of each cluster, correctly resulting in the archetypes being extreme observations and centroids being median representations of

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

Fig. 7. Radar plot of the centroids obtained in the ALS case study.

the cluster's points. This leads to values more markedly towards the extremes of the ranges in the original variable space for the archetypes than for the centroids. Moving on with the analysis in an attempt to characterise the disease states represented by the archetypes or centroids respectively, it is fairly straightforward to associate each archetype with the corresponding impaired domain(s) by observing the radar plots. For the centroids, although the shape on the radar plot is on the whole preserved, the fact that the edges are more moderate makes identifying and interpreting a precise disease state less clear. By comparing the way the sequence of visits for each subject is mapped onto the disease states obtained as archetypes or centroids, different scenarios can be observed, with the transitions between the states encoded with the centroids being either corresponding, earlier, or later than those between the archetypes. This is expected, as they are essentially complementary descriptions of how the pathology evolves in subjects, namely sequences of extreme states (with archetypes) or median states (with centroids). Figure 8 reports the PCA plot of the observed visits, including archetypes and centroids. This plot aligns with previous observations, showing archetypes positioned farther outward than the centroids, and showing a correspondence between some archetypes and centroids (e.g., archetype "none" and centroid "C4", or archetype "LL" and centroid "C5"), with centroids consistently placed in more central positions relative to their corresponding archetypes.

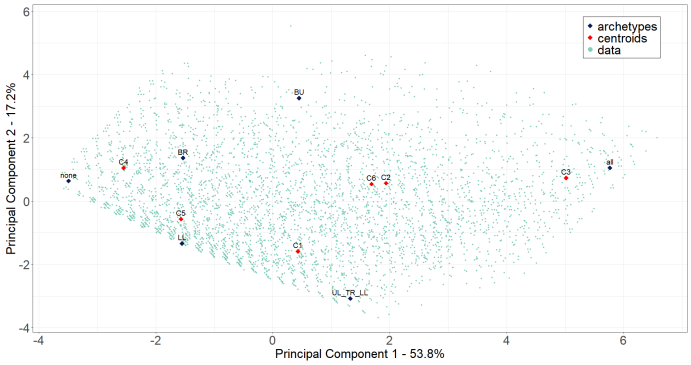

Fig. 8. PCA of ALS data, derived archetypes and centroids.

In summary, this comparison of archetypes and clustering techniques to define disease states indicates that archetypes are effectively able to provide more extreme representations of the condition of subjects, compared to traditional clustering methods, such as k-means, which provide average and less definite descriptions of the disease states. In certain contexts, such as that of the case study presented in this work, the use of AA as proposed in DYNAMITE may be preferable to describe the course of pathologies due to its enhanced interpretability. In other cases, it might be of interest to implement both analyses, given their complementarity. Furthermore, both approaches can be considered alternative and complementary to methods using PM directly in clinically obtained descriptive states, such as proposed in [31] where the more compact Milano-Torino Staging System (MiToS) [38] was used to describe the progression of ALS as a cumulative sequence of functional impairments.

It is worth noting that while the presented case study focuses on a disease of a progressively degenerative nature and uses an established clinical scale to quantify the increase in disability, the proposed methodology could be used in general for applications with any pattern of variability. Starting from a description of the patient's condition over time, with the use of scales (as in our case study) or any other clinically useful descriptor (such as haematochemical parameters), the use of AA can help define, according to their definition, extreme states representative of the clinical condition in the population. In the case of conditions that are not necessarily characterised by progressive worsening over time, such states may, for instance, correspond to phases of improvement or worsening, or represent extreme intermediate phases. In the latter case, such states may be more complex to map in clinical evidence than in our case study, where it was quite straightforward to map them to the functional domains affected by ALS. Continuing in the application of our methodology, in the subsequent definition of trajectories by PM, what is expected in the case of clinical conditions characterised by a wider irregularity is that the identified patterns may be greater in number (if the condition presents great heterogeneity in the population) or that oscillations between states may be observed (where the condition presents alternating phases of improvement or worsening).

In studying clinical conditions where the population of interest includes both rapid and slow progressors (as in the presented ALS case study), moreover, one must also take into account that the proposed methodology encapsulates these two varying patterns collectively. In these cases, it is appropriate to make some specific considerations. For slow progressors, it is necessary to evaluate whether the data characteristics (especially in terms of follow-up length relative to the expected prognosis) allow for the observation of significant changes in disease states, or if there is a risk that subjects might be censored without ever changing state during their consecutive visits. To address this, it may be useful to conduct additional analyses for this population, such as examining the duration within the same disease state, or to extend the clinical observation period if possible. With reference to fast progressors, the risk lies in not having sufficiently frequent sampling to accurately track all state changes. Although it is true that this would result in the loss of interim archetypes in our methodology, this primarily reflects what could happen in clinical practice, where clinicians may miss some steps of disease progression if they do not have the opportunity to visit subjects within adequate time intervals. In our case study, we believe that the sampling was sufficiently frequent and the follow-up period was adequately long (also thanks to the fact that we used real-world data rather than clinical trial data) to assume that we have accurately described both subpopulations.

A limitation of the proposed methodology is that, in its current implementation, only numeric variables are used as input. However, longitudinal clinical datasets often contain mixed-type variables, which may require conversion to dummy variables for use within DYNAMITE. Although this preprocessing step is feasible, it may be suboptimal since AA was designed to work with continuous features. Another aspect concerns the fact that in the present version of our work, DYNAMITE exclusively employs dynamic data, overlooking

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

the potential contributions of static features (e.g. demographics) in characterising disease states and patient trajectories. However, additional analyses may complement the proposed methodology to fully exploit available information. For example, one could explore the integration of these features in the analysis by appending the static variables to each patient's visit, or, alternatively, by including them *a posteriori* for stratifying and examining trajectories for different groups of patients.

Kaplan-Meier curves can be used to inspect differences in progression rates among different subgroups. As an example, in our case study, we extracted from the whole population the subjects whose first visit was associated through our methodology with a state of low disability (corresponding to the "none" archetype) and for whom the "Dead" event was later recorded. This resulted in the selection of 407 subjects. On this sub-population, we separated subjects presenting a bulbar impairment as the first impacted domain (i.e., the subjects with their following disease status corresponding to the "BU" archetype) from those presenting motor symptoms (i.e., the subjects with their following disease status corresponding to the "LL" or "UL_TR_LL" archetypes). We then computed the Kaplan-Meier curves (reported in Figure 9) to estimate and compare the survival functions of the two populations.

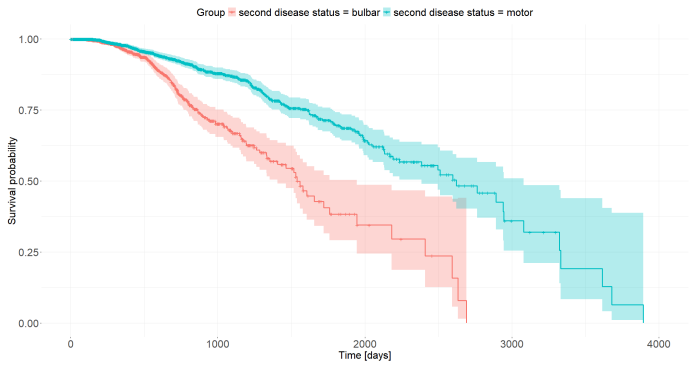

Fig. 9. Kaplan-Meier survival curves of the population of subjects entering the study in a state of low disability and for whom the death event is registered, stratified according to their second identified disease status: subjects associated with a bulbar domain impairment (in red) versus subjects associated with an impairment in the motor domain (in blue).

This analysis effectively highlighted slow versus rapid progressors: the group with early bulbar impairment has a median time to death of 656 days (IQR 420-746), which increases to 829 days (IQR 449-1350) for those with early motor impairments. This is consistent with the medical literature, where subjects with bulbar onset are consistently reported to have a worse outcome [39].

Future directions include extending DYNAMITE by investigating alternative rules for associating each observation with an archetype in a fuzzy manner. This exploration could be particularly valuable for instances where a visit is located in the center of the space defined by the archetypes, resulting in an observation that is similar to multiple archetypes. Additionally, we intend to integrate static and mixed-type features to further extend and automate the analysis capacity, enhancing the characterisation of disease states in patients with clinical conditions and accurately depicting their progression paths.

Code Availability

The code implementing the method proposed in this manuscript is available, together with a synthetic dataset simulating a longitudinal dataset like the one used for the case study in the manuscript, at this link.

Author Contribution

BDC and MV designed and supervised the study; IT, ET, MV, and BDC developed the methodology, IT, ET, and RG performed the case-study analyses; IT, ET, and BDC wrote the paper; RV and AC provided the case-study data; all authors contributed to the discussion and were involved in critically revising the manuscript and approving the final version to be submitted.

References

[1] I. Ramos-Vielba, N. Robinson-Garcia, and R. Woolley, "A value creation model from science-society interconnections: Archetypal analysis combining publications, survey and altmetric data," *Plos one*, vol. 17, no. 6, p. e0269004, 2022.

[2] Z. Wang, Y. Lu, G. Zhao, C. Sun, F. Zhang, and S. He, "Sugarcane biomass prediction with multi-mode remote sensing data using deep archetypal analysis and integrated learning," *Remote Sensing*, vol. 14, no. 19, p. 4944, 2022.

[3] U. Grzybowska and M. Karwański, "Archetypal analysis and dea model, their application on financial data and visualization with phate," *Entropy*, vol. 24, no. 1, p. 88, 2022.

[4] J. I. Stoker, H. Garretsen, D. Soudis, and T. Vriend, "A configurational approach to leadership behavior through archetypal analysis," *Frontiers in Psychology*, vol. 13, p. 1022299, 2023.

[5] P. Vicente and A. Suleman, "Covid-19 in europe: from outbreak to vaccination," *BMC public health*, vol. 22, no. 1, pp. 1–17, 2022.

[6] S. M. Groves, G. V. Ildefonso, C. O. McAtee, P. M. Ozawa, A. S. Ireland, P. E. Stauffer, P. T. Wasdin, X. Huang, Y. Qiao, J. S. Lim, *et al.*, "Archetype tasks link intratumoral heterogeneity to plasticity and cancer hallmarks in small cell lung cancer," *Cell Systems*, vol. 13, no. 9, pp. 690–710, 2022.

[7] J. Reeve, G. A. Böhmig, F. Eskandary, G. Einecke, C. Lefaucheur, A. Loupy, P. F. Halloran, M.-K. S. Group, *et al.*, "Assessing rejection-related disease in kidney transplant biopsies based on archetypal analysis of molecular phenotypes," *JCI insight*, vol. 2, no. 12, 2017.

[8] P. Hruba, J. Klema, A. V. Le, E. Girmanova, P. Mrazova, A. Massart, D. Maixnerova, L. Voska, G. B. Piredda, L. Biancone, *et al.*, "Novel transcriptomic signatures associated with premature kidney allograft failure," *EBioMedicine*, vol. 96, 2023.

[9] J. Gimbernat-Mayol, A. Dominguez Mantes, C. D. Bustamante, D. Mas Montserrat, and A. G. Ioannidis, "Archetypal analysis for population genetics," *PLoS Computational Biology*, vol. 18, no. 8, p. e1010301, 2022.

[10] C. Pancotti, C. Rollo, G. Birolo, S. Benevenuta, P. Fariselli, and T. Sanavia, "Unravelling the instability of mutational signatures extraction via archetypal analysis," *Frontiers in Genetics*, vol. 13, p. 1049501, 2023.

[11] S. Yousefi, L. R. Pasquale, M. V. Boland, and C. A. Johnson, "Machine-identified patterns of visual field loss and an association with rapid progression in the ocular hypertension treatment study," *Ophthalmology*, vol. 129, no. 12, pp. 1402–1411, 2022.

[12] T. Elze, L. R. Pasquale, L. Q. Shen, T. C. Chen, J. L. Wiggs, and P. J. Bex, "Patterns of functional vision loss in glaucoma determined with archetypal analysis," *Journal of The Royal Society Interface*, vol. 12, no. 103, p. 20141118, 2015.

[13] J. Branco, T. Elze, J.-K. Wang, L. R. Pasquale, M. K. Garvin, R. Kardon, and M. J. Kupersmith, "Longitudinal visual field archetypal analysis of optic neuritis treated in a clinical setting," *BMJ open ophthalmology*, vol. 7, no. 1, p. e001136, 2022.

[14] E. Solli, H. Doshi, T. Elze, L. R. Pasquale, J. Branco, M. Wall, and M. Kupersmith, "Archetypal analysis of visual fields in optic neuritis reveals functional biomarkers associated with outcome and treatment response," *Multiple Sclerosis and Related Disorders*, vol. 67, p. 104074, 2022.

[15] H. Doshi, E. Solli, T. Elze, L. R. Pasquale, M. Wall, and M. J. Kupersmith, "Unsupervised machine learning identifies quantifiable patterns of visual field loss in idiopathic intracranial hypertension," *Translational Vision Science & Technology*, vol. 10, no. 9, pp. 37–37, 2021.

[16] M. Eugster and F. Leisch, "From spider-man to hero-archetypal analysis in r," 2009.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3453602

[17] I. Trescato, E. Tavazzi, M. Vettoretti, R. Vasta, A. Chiò, and B. Di Camillo, "Identifying extreme profiles in amyotrophic lateral sclerosis patients at diagnosis through archetypal analysis," 2023. Cited by: 0.

[18] R. Bellazzi, F. Ferrazzi, and L. Sacchi, "Predictive data mining in clinical medicine: a focus on selected methods and applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 5, pp. 416–430, 2011.

[19] E. De Roock and N. Martin, "Process mining in healthcare–an updated perspective on the state of the art," *Journal of biomedical informatics*, p. 103995, 2022.

[20] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *Journal of biomedical informatics*, vol. 61, pp. 224–236, 2016.

[21] W. M. van der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. van Den Brand, R. Brandtjen, J. Buijs, *et al.*, "Process mining manifesto," in *International Conference on Business Process Management*, pp. 169–194, Springer, 2011.

[22] W. M. van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1128–1142, 2004.

[23] W. M. van der Aalst, A. Adriansyah, and B. van Dongen, "Replaying history on process models for conformance checking and performance analysis," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 182–192, 2012.

[24] W. M. van der Aalst, *Process mining: discovery, conformance and enhancement of business processes*, vol. 2. Springer, 2011.

[25] A. Cutler and L. Breiman, "Archetypal analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[26] M. J. Eugster and F. Leisch, "Weighted and robust archetypal analysis," *Computational Statistics & Data Analysis*, vol. 55, no. 3, pp. 1215–1225, 2011.

[27] G. Ragozini, F. Palumbo, and M. R. D'Esposito, "Archetypal analysis for data-driven prototype identification," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 10, no. 1, pp. 6–20, 2017.

[28] A. Dagliati, L. Sacchi, A. Zambelli, V. Tibollo, L. Pavesi, J. H. Holmes, and R. Bellazzi, "Temporal electronic phenotyping by mining careflows of breast cancer patients," *Journal of biomedical informatics*, vol. 66, pp. 136–147, 2017.

[29] A. Dagliati, V. Tibollo, G. Cogni, L. Chiovato, R. Bellazzi, and L. Sacchi, "Careflow mining techniques to explore type 2 diabetes evolution," *Journal of diabetes science and technology*, vol. 12, no. 2, pp. 251–259, 2018.

[30] M. A. Cuendet, R. Gatta, A. Wicky, C. L. Gerard, M. Dalla-Vale, E. Tavazzi, G. Michielin, J. Delyon, N. Ferahta, J. Cesbron, *et al.*, "A differential process mining analysis of covid-19 management for cancer patients," *Frontiers in Oncology*, vol. 12, p. 1043675, 2022.

[31] E. Tavazzi, R. Gatta, M. Vallati, S. Cotti Piccinelli, M. Filosto, A. Padovani, M. Castellano, and B. Di Camillo, "Leveraging process mining for modeling progression trajectories in amyotrophic lateral sclerosis," *BMC Medical Informatics and Decision Making*, vol. 22, no. 6, pp. 1–17, 2022.

[32] R. Gatta, M. Vallati, J. Lenkowicz, E. Rojas, A. Damiani, L. Sacchi, B. De Bari, A. Dagliati, C. Fernandez-Llatas, M. Montesi, *et al.*, "Generating and comparing knowledge graphs of medical processes using pminer," in *Proceedings of the Knowledge Capture Conference*, pp. 1–4, 2017.

[33] A. Chiò, G. Mora, C. Moglia, U. Manera, A. Canosa, S. Cammarosano, A. Ilardi, D. Bertuzzo, E. Bersano, P. Cugnasco, M. Grassano, F. Pisano, L. Mazzini, A. Calvo, for the Piemonte, and V. d'Aosta Register for ALS (PARALS), "Secular Trends of Amyotrophic Lateral Sclerosis: The Piemonte and Valle d'Aosta Register," *JAMA Neurology*, vol. 74, pp. 1097–1104, 09 2017.

[34] B. Brooks, M. Sanjak, S. Ringel, J. England, and J. Brinkmann, "ym.(1996). the amyotrophic lateral sclerosis functional rating scale: Assessment of activities of daily living in patients with amyotrophic lateral sclerosis," *Archives of Neurology*, vol. 53, no. 2, pp. 141–147.

[35] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi, "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the Neurological Sciences*, vol. 169, no. 1, pp. 13 – 21, 1999.

[36] B. R. Brooks, R. G. Miller, M. Swash, and T. L. Munsat, "El escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotrophic lateral sclerosis and other motor neuron disorders*, vol. 1, no. 5, pp. 293–299, 2000.

[37] A. Chiò, A. Calvo, C. Moglia, L. Mazzini, G. Mora, *et al.*, "Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study," *Journal of Neurology, Neurosurgery & Psychiatry*, pp. jnnp–2010, 2011.

[38] A. Chiò, E. R. Hammond, G. Mora, V. Bonito, and G. Filippini, "Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 86, no. 1, pp. 38–44, 2015.

[39] A. Chio, G. Logroscino, O. Hardiman, R. Swingler, D. Mitchell, E. Beghi, B. G. Traynor, E. Consortium, *et al.*, "Prognostic factors in als: a critical review," *Amyotrophic Lateral Sclerosis*, vol. 10, no. 5-6, pp. 310–323, 2009.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/