scientific reports

Check for updates

OPEN

Identifying time patterns in Huntington's disease trajectories using dynamic time warping-based clustering on multi-modal data

Alexia Giannoula[1]✉, Audrey E. De Paepe[1,2], Ferran Sanz[1], Laura I. Furlong[3] & Estela Camara[2]

One of the principal goals of Precision Medicine is to stratify patients by accounting for individual variability. However, extracting meaningful information from Real-World Data, such as Electronic Health Records, still remains challenging due to methodological and computational issues. A Dynamic Time Warping-based unsupervised-clustering methodology is presented in this paper for the clustering of patient trajectories of multi-modal health data on the basis of shared temporal characteristics. Building on an earlier methodology, a new dimension of time-varying clinical and imaging features is incorporated, through an adapted cost-minimization algorithm for clustering on different, possibly overlapping, feature subsets. The model disease chosen is Huntington's disease (HD), characterized by progressive neurodegeneration. From a wide range of examined user-defined parameters, four case examples are highlighted to demonstrate the identified temporal patterns in multi-modal HD trajectories and to study how these differ due to the combined effects of feature weights and granularity threshold. For each identified cluster, polynomial fits that describe the time behavior of the assessed features are provided for an informative comparison, together with their averaged values. The proposed data-mining methodology permits the stratification of distinct time patterns of multi-modal health data in individuals that share a diagnosis, by employing user-customized criteria beyond the current clinical practice. Overall, this work bears implications for better analysis of individual variability in disease progression, opening doors to personalized preventative, diagnostic and therapeutic strategies.

Keywords Precision medicine, Longitudinal cohort analysis, Multi-modal Real-World Data, Patient stratification, Unsupervised clustering, Time analysis

Precision medicine is an innovative approach that takes into account individual variability in genes, environment and lifestyle to understand diseases at a multifactorial level, with the aim of optimizing prevention, diagnostics and treatment of patient subgroups[1–3]. Such work has been empowered by the rapid expansion of Real-World Data (RWD), such as data stored within electronic healthcare records (EHRs), making precision medicine increasingly popular in recent years. In this context, individuals with a common diagnostic label may exhibit differences in terms of disease onset and response to treatment, characteristics that necessitate identification to maximize patient wellbeing. However, due to the challenges posed in extracting useful and actionable information from the large volumes of available data, there is an urgent need for the development of novel methodological approaches and computational tools to effectively stratify patients based on their relevant individual differences.

In this respect, the use of data-mining technology has introduced new prospects in the field of biomedical research[4–7], as it permits extracting a wealth of knowledge from health datasets that otherwise would remain hidden within patients' clinical histories. In light of this, a growing number of studies has employed different algorithms to explore EHRs that mainly contain common clinical descriptors, such as diagnostic codes, human phenotypes, lab tests and other clinical features[8–12]. However, in the majority of these works, the temporal dimension is typically not taken into account or is considered in an insufficient way. It is only in the recent years that patterns of disease progression, referred to as trajectories, have been studied with the use of longitudinal

[1]Research Group on Integrative Biomedical Informatics (GRIB), Department of Medicine and Life Sciences (MELIS), Universitat Pompeu Fabra, Hospital del Mar Research Institute, Barcelona, Spain. [2]Cognition and Brain Plasticity Unit, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Barcelona, Spain. [3]MedBioinformatics Solutions, Barcelona, Spain. ✉email: alexia.giannoula@upf.edu

patient cohorts[13–18]. These studies aimed to identify subgroups of patients exhibiting similar progression patterns by means of data-mining techniques for a variety of disease areas, including heart failure, diabetes mellitus or prostate cancer. This was achieved by monitoring the temporal dynamics of specific health outcomes, such as disease state or diagnostic codes.

In this study, we present an innovative data-mining methodological framework for the identification of distinct patient groups by analyzing the progression of a broad set of clinical and imaging features over time. This is achieved by significantly expanding and adapting the Dynamic Time Warping (DTW)-based unsupervised clustering methodology implemented in prior works[15,18,19], while incorporating a cost-minimization algorithm from[20]. The latter enables clustering on different and possibly overlapping subsets of the aforementioned features. Clustering trajectories based on a series of time-varying features constitutes an important contribution to the -so far- limited literature of data-mining techniques for longitudinal cohort analyses in the field of biomedicine.

To demonstrate the potential of the proposed methodology in elucidating the heterogeneity in patients' clinical profiles, we employ Huntington's disease (HD) as a model disease[21]. In brief, HD is an inherited neurodegenerative disease caused by a CAG repeat expansion in the *HTT* gene. Neurodegeneration in HD patients is progressive and debilitating. Brain atrophy initially targets the striatum, a deep brain structure whose subregions include the caudate, putamen and nucleus accumbens[22–24]. The disease typically manifests in mid-adulthood with a characteristic triad of motor, cognitive and psychiatric features[25–27]. Gene-expansion carriers who will later develop the disease (e.g., premanifest individuals) can be identified prior to overt disease onset by elective genetic testing. Considering this and given its progressive nature, HD has previously been described as a model for neurodegeneration[28]. At the same time, HD is heterogeneous in clinical presentation and disease course, highlighting its potential for improved descriptions of disease progress and tailored therapeutic interventions. As such, given the variability in disease progression and the ability to identify susceptible individuals prior to overt disease onset, HD constitutes a suitable model to showcase the potential of the proposed methodology in revealing subgroups of patients with similar profiles of disease progression. Clinical evaluations describing the evolution of clinical features in all three domains (motor, cognitive and psychiatric) are considered, along with three target brain volumes derived from Magnetic Resonance Imaging (MRI): putamen, caudate and nucleus accumbens.

Several traditional techniques exist in the literature that analyze static features. For example, in[29] a machine-learning technique was presented to identify genes contributing to HD from gene-expression profiles, while in[30], unsupervised clustering was applied on speech assessments to identify motor speech patterns. In both works, time changes of features were ignored, and a single data type was considered (e.g. speech assessments or gene profiles), in contrast to the current method that explores time dynamics across various data modalities.

In the context of longitudinal analyses, unified frameworks to subtype disease progression profiles have recently been executed in neurodegenerative disease, such as in Alzheimer's disease, amyotrophic lateral sclerosis and Parkinson's disease[31,32]. In HD specifically, recent examples of trajectory analyses can be found in[33–36]. Most of these works involve computationally intensive processes (e.g. deep learning, Gaussian/Dirichlet) that require large datasets to be trained and may also hinder direct interpretability of results, by obscuring the understanding of what specific patterns or features are driving the clusters. Furthermore, performing the analysis on a single composite score or a small subset of features may compromise the ability to understand the progression of disease globally, by missing subtle variations in specific disease domains.

On the other hand, the proposed methodology provides a more detailed, flexible, and holistic approach to the study of Huntington's disease progression. By directly aligning time sequences of diverse data modalities, both from the clinical and neuroimaging space, the DTW-based clustering technique captures intricate temporal patterns and variations in each disease domain. Overall, this study showcases the potential of the methodology in longitudinal patient stratification, by also allowing customization of the clustering parameters to meet specific clinical needs. Its findings are expected to serve as a preliminary basis for better understanding the variability in disease profiles, in this instance exemplified in individuals with HD. Ultimately, the proposed methodology can be appropriately adapted and applied to any disease of interest and observational cohort, for which multidimensional metrics are available in the time axis.

Results

The proposed DTW-based unsupervised algorithm was applied on the HD trajectories of $n = 44$ individuals using the following values for the feature contribution parameter λ: 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. With respect to the granularity threshold $thres_{gr}$, a range of values between 0.5 and 8.0 was considered with intervals of 0.25. For each combination of $thres_{gr}$ and λ, the *averaged total distance* over all identified clusters was calculated as previously described along with the number of extracted clusters (Supplementary Fig. S1). Finally, nine pairs $(thres_{gr}, \lambda)$ were found to be optimal by fulfilling the requirements set in Methods, i.e.: (1.25, 0.01), (1.50, 0.01), (1.75, 0.01), (2.0, 0.1), (2.25, 0.01), (2.25, 0.1), (2.5, 0.01), (2.75, 0.01), (3.0, 0.01). Four case examples are highlighted below to demonstrate the effects of the aforementioned parameters on the identified clusters.

Case 1: $thres_{gr} = 1.50$ and $\lambda = 0.01$

Case 1 demonstrated one of the optimal combinations that satisfied the conditions defined in the cluster evaluation process (see "Methods"), specifically $thres_{gr} = 1.50$ and $\lambda = 0.01$. This resulted in a total of ten identified clusters as depicted in Fig. 1, with a total average distance of 0.66. Table 1 details averaged feature values, as well as averaged fitted polynomials and distribution of the final score weights (i.e., those reached at the final step of the iterative clustering algorithm). In the majority of clusters, only one or at most two features contributed to clustering, as was expected due to the very low selected value of λ. The square of the Pearson correlation coefficient (R2) is also reported for the averaged fitted polynomials of each cluster in Fig. 1 (values

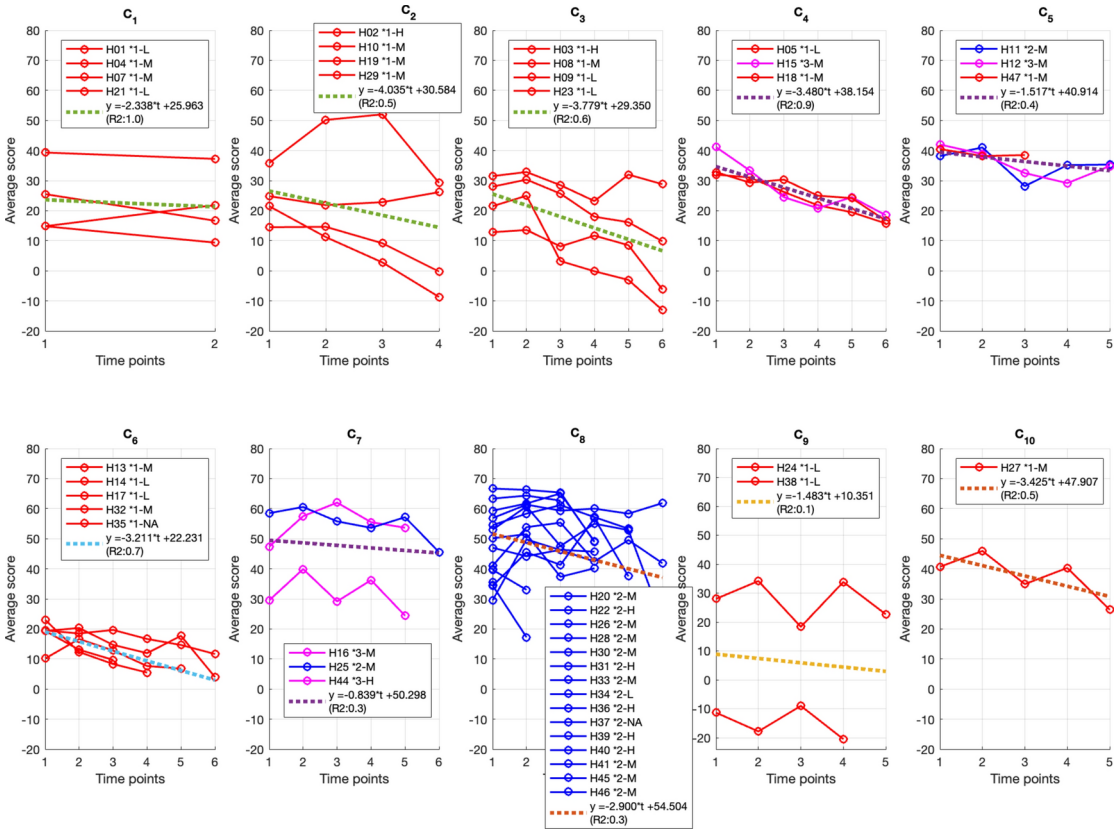

Fig. 1. Extracted clusters of HD trajectories for Case 1 ($thres_{gr} = 1.50$ and λ = 0.01), versus ordered time points. Red depicts manifest profiles, blue premanifest and pink those individuals that transitioned from premanifest to manifest over the course of assessment. The average score on the y-axis is produced by averaging all six features at each time point. The fitted polynomial of degree 1 is also shown (dotted lines) in each cluster averaged over all patients, together with the R2 (square of the Pearson correlation coefficient) *. *In C_1, all four patients have two data points and thus, the fitted line passes through both points (R2 = 1.0).

closer to 1.0 indicate better fitting). Individual Silhouette scores and average-distance values per cluster can be seen in Supplementary Table S1.

A very clear distinction of individuals in the premanifest phase was observed, where the majority was assigned to cluster C_8 based exclusively on the motor feature ($k = 2$). This cluster was distinguished by the presence of a very low level of motor symptoms, as evidenced by the fitted polynomial $y_{\text{fit},2}$ (Table 1). Furthermore, individuals in C_8 (together with C_7) presented the mildest cognitive and motor symptoms. In addition, the extent of brain atrophy in the caudate and putamen, represented by the MRI features, was the least in C_8 compared to all ten clusters.

Conversely, manifest patients were further subdivided into distinct clusters according to the time characteristics of the corresponding associated feature vectors. Notably, two major subgroups (clusters C_1 and C_2) of manifest patients (group 1) were formed, clustered based on the atrophy of the MRI putamen volume ($k = 5$). Patients assigned to C_2 demonstrated lower severity in all clinical scores and brain volumes (except psychiatric) at enrollment but deteriorated significantly faster than those in C_1 Clusters C_4-C_6 were formed with the exclusive participation of the total cognitive score ($k = 1$) and specifically, patients in C_4 exhibited one of the fastest rates of cognitive decline ($y_{fit,1}$).

Case 2: $thres_{gr} = 3.75$ and λ = 0.01

In Case 2, λ was kept consistent with Case 1 (λ = 0.01), but a higher granularity threshold was selected ($thres_{gr} = 3.75$). Although these parameter values did not align with the list of empirically optimal combinations (see beginning of the "Results" section), they were selected to illustrate how certain parameters can result in a smaller number of extracted clusters, that is, coarser clustering at the cost of reduced homogeneity (see Methods). As such, the clustering algorithm converged into four clusters (Fig. 2), with an ensuing increase in total average distance (1.87 vs 0.66 in Case 1), as expected. Table 2 details averaged feature values, fitted polynomials and score weights. Silhouette scores and average-distance values per cluster can be found in Supplementary Table S1.

Similar to Case 1, Case 2 resulted again in one predominant premanifest cluster (C_2), this time with three principally manifest clusters (C_1, C_3, C_4). Specifically, the premanifest cluster C_2 (similar to C_8 in Case 1), was characterized by very low motor deficits with a relatively slow evolution over time (see $y_{fit,2}$ in Table 2). On

	C_l	1	2	3	4	5	6	7	8	9	10
Averaged feature scores	Cogn. sc (ω_{1l})	171.74 (0.0)	169.2 (0.0)	145.0 (0.0)	188.8 (1.0)*	242.7 (1.0)*	135.3 (0.99)*	306.3 (0.0)	299.4 (0.0)	128.8 (0.0)	237.8 (0.17)*
	Mot. sc (ω_{2l})	-24.7 (0.0)	-26.6 (0.0)	-33.8 (0.0)	-17.2 (0.0)	-11.7 (0.0)	-42.2 (0.0)	-3.2 (0.92)*	-0.6 (1.0)*	-44.0 (0.0)	-8.6 (0.17)*
	Psych. sc (ω_{3l})	-12.3 (0.0)	-19.6 (0.0)	-19.5 (0.0)	-15.8 (0.0)	-10.8 (0.0)	-10.8 (0.01)	-18.6 (0.0)	-14.0 (0.0)	-46.6 (0.0)	-3.4 (0.17)*
	Caud. vol (ω_{4l})	1.1e-3 (0.0)	1.3e-3 (0.0)	1.0e-3 (0.0)	1.3e-3 (0.0)	1.6e-3 (0.0)	1.0e-3 (0.0)	1.5e-3 (0.0)	1.7e-3 (0.0)	9.3e-4 (1.0)*	1.7e-3 (0.17)*
	Put. vol (ω_{5l})	1.5e-3 (1.0)*	1.8e-3 (1.0)*	1.3e-3 (0.47)*	1.7e-3 (0.0)	2.0e-3 (0.0)	1.4e-3 (0.0)	2.0e-3 (0.08)*	2.2e-3 (0.0)	1.2e-3 (0.0)	2.2e-3 (0.17)*
	Acc. vol (ω_{16})	1.4e-4 (0.0)	1.6e-4 (0.0)	1.2e-4 (0.53)*	1.5e-4 (0.0)	1.7e-4 (0.0)	1.3e-4 (0.0)	2.2e-4 (0.0)	2.1e-4 (0.0)	1.1e-4 (0.0)	1.7e-4 (0.17)*
Averaged fitted polynomials of degree 1	Cogn. sc. $y_{\text{fit},1}$	-10.5*t+187.5	-17.0*t+211.7	-15.6*t+204.7	-16.2*t+245.5	-8.3*t+265.3	-13.9*t+172.2	-7.9*t+331.8	-15.7*t+290.6	-4.7*t+141.2	-19.4*t+296.1
	Mot. sc $y_{\text{fit},2}$	1.1*t -26.4	-9.9*t -2.0	-5.5*t -14.6	-3.1*t -6.4	-2.2*t -5.3	-6.5*t -26.0	-1.5*t +1.6	-0.4*t +0.2	-4.5*t -31.8	-0.5*t -7.1
	Psych. sc $y_{\text{fit},3}$	-4.6*t -5.4	2.7*t -26.2	-1.6*t -13.9	-1.6*t -10.1	1.4*t -14.5	1.1*t -12.8	4.4*t -31.6	-1.3*t -9.1	0.3*t -47.2	-0.6*t -1.6
	Caud. vol $y_{\text{fit},4}$	0.6e-4*t+1.0e-3	-3.1e-4*t+2.1e-3	-2.1e-4*t+1.7e-3	-2.5e-4*t+2.2e-3	-3.6e-4*t+2.6e-3	-2.7e-4*t+1.8e-3	-3.1e-4*t+2.5e-3	-3.8e-4*t+2.4e-3	-2.9e-4*t+1.7e-3	-3.1e-4*t+2.6e-3
	Put. vol $y_{\text{fit},5}$	-0.9e-4*t+1.6e-3	-4.3e-4*t+2.9e-3	-2.3e-4*t+2.1e-3	-3.4e-4*t+2.8e-3	-4.6e-4*t+3.2e-3	-3.3e-4*t+2.4e-3	-4.1e-4*t+3.3e-3	-4.6e-4*t+2.9e-3	-3.6e-4*t+2.2e-3	-3.2e-4*t+3.1e-3
	Acc. vol. $y_{\text{fit},6}$	-0.9e-5*t+1.5e-4	-4.7e-5*t+2.8e-4	-2.2e-5*t+1.9e-4	-3.2e-5*t+2.6e-4	-3.8e-5*t+2.7e-4	-3.0e-5*t+2.1e-4	-4.7e-5*t+3.7e-4	-4.0e-5*t+2.7e-4	-3.5e-5*t+2.1e-4	-4.5e-5*t+3.0e-4
	Averaged $y_{\text{fit,all}}$	-2.3*t+26.0	-4.0*t+30.6	-3.8*t+29.4	-3.5*t+38.2	-1.5*t+40.9	-3.2*t+22.2	-0.8*t+50.3	-2.9*t+54.5	-1.5*t+10.4	-3.4*t+47.9

Table 1. Case 1 averaged feature scores over all patients and times for each cluster (C_l l=1–10), for $thres_{gr} = 1.50$ and λ = 0.01. The corresponding distribution of the score weights $\omega_{kl}, k = 1-6$ is also reported in parenthesis. Subsequently, the averaged fitted polynomials of degree 1 are shown with respect to time averaged over all patients for each feature separately ($y_{\text{fit},k}$, k=1–6) and averaged over all features and all patients ($y_{\text{fit,all}}$). Higher score values indicate improved functioning. *The asterisk denotes feature scores that contributed more than 5% to the cluster assignment (i.e.$\omega_{kl} > 0.05$); The abbreviated descriptions of the left column refer to the following features (sequentially, see also Table S2): total cognitive score, total motor score, total psychiatric score, caudate volume, putamen volume and nucleus accumbens volume (the three latter denote averaged MRI volumes across right and left hemispheres for each participant).

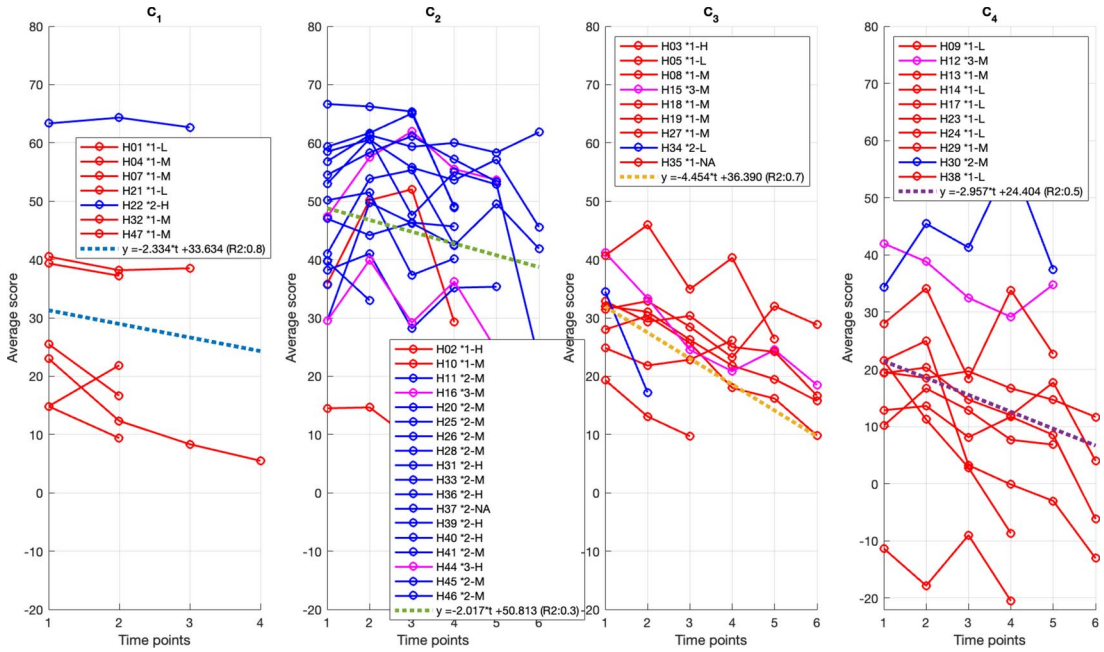

Fig. 2. Extracted clusters of HD trajectories for Case 2 ($thres_{gr} = 3.75$ and λ = 0.01), versus ordered time points. Red depicts manifest profiles, blue premanifest and pink those individuals that transitioned from premanifest to manifest over the course of assessment. The average score on the y-axis is produced by averaging all six features at each time point. The fitted polynomial of degree 1 is also shown (dotted lines) in each cluster averaged over all patients, together with the R2 (square of the Pearson correlation coefficient).

the other hand, the manifest cluster C_1 was marked by the best (i.e., least severe) averaged total psychiatric symptoms but with the fastest deterioration in slope ($y_{fit,3}$) of the psychiatric feature. Meanwhile, the manifest cluster C_3 included mainly manifest patients undergoing the fastest decline of the total cognitive score ($y_{fit,1}$) of all manifest clusters. Finally, C_4 was formed based on brain atrophy of both the caudate and accumbens volumes, encompassing patients with the fastest decline in the total motor score ($y_{fit,2}$) with stable evolution of the psychiatric feature ($y_{fit,3}$) over time.

Case 3: $thres_{gr} = 3.75$ and λ = 0.4

Case 3 exemplifies the effect of λ when patients are selectively grouped by a larger number of features. For this purpose, the granularity threshold was kept identical as in Case 2 ($thres_{gr} = 3.75$) and λ was set equal to 0.4 (this combination of parameters is not listed as empirically optimal in the beginning of Results). Increasing λ permits a higher number of features to take part in the formation of each cluster. As a result, eight clusters were extracted (Fig. 3), and averaged feature values, fitted polynomials and score weights were produced (Table 3). The total average distance increased to 2.03 compared to Case 2, indicating a decrease in cluster homogeneity, as expected, since forming clusters with small dispersion in all or many features is not a practically optimal scenario (see "Methods").

In Case 3, increasing the feature contribution parameter led to splitting of the major premanifest cluster of Case 2 (C_2 in Fig. 2) into several smaller mainly premanifest or mixed clusters (e.g. C_1, C_5, C_6, C_7 in Fig. 3) and the rearrangement of the three principal manifest clusters of Case 2. In particular, the two purely premanifest clusters C_5 and C_7 contained the least severe cases of all with similar evolution over time in basically all clinical and brain-volume features except for the psychiatric domain. In this respect, patients in C_5 exhibited a significantly worse psychiatric score than C_7 at enrollment that improved relatively quickly with time, while patients in C_7 remained relatively stable at higher (less severe) levels ($y_{fit,3}$ in Table 3). At the same time, Case 3 demonstrated two purely manifest clusters (C_3 and C_4), where manifest patients in C_4 were overall more severe than C_3 and deteriorated faster in all clinical and brain-volume features, except for the psychiatric domain that was stable through time ($y_{fit,3}$).

Case 4: $thres_{gr} = 6.0$ and λ = 0.4

In Case 4, the granularity threshold $thres_{gr}$ was further increased to 6.0, while λ was maintained at 0.4 (this combination of parameters is also not listed as empirically optimal in the beginning of Results). This resulted in the coarsest possible separation, dividing patients into a manifest (C_1) and premanifest (C_2) cluster (Fig. 4). The resulting total average distance was the highest compared to all previous case examples, reaching 3.75. Furthermore, although all six features contributed to the formation of the two clusters (see score weights ω_{kl} in Table 4), the most notable differences were seen in the total motor deficits ($k = 2$), which was significantly lower and deteriorated at a slower rate in C_2 than C_1. Interestingly, the decline in the total cognitive score was similar in both clusters ($y_{\text{fit},1}$), and the brain atrophy represented by the three MRI features was faster in premanifest

	C_l	1	2	3	4
Averaged feature scores	Cogn. sc (ω_{1l})	207.5 (0.0)	287.8 (0.0)	191.0 (1.0)*	148.9 (0.0)
	Mot. sc (ω_{2l})	-22.1 (0.0)	-4.1 (1.0)*	-20.5 (0.0)	-34.2 (0.0)
	Psych. sc (ω_{3l})	-10.0 (0.99)*	-14.1 (0.0)	-17.1 (0.0)	-22.3 (0.0)
	Caud. vol (ω_{4l})	1.3e-3 (0.0)	1.7e-3 (0.0)	1.2e-3 (0.0)	1.1e-3 (0.33)*
	Put. vol (ω_{5l})	1.7e-3 (0.0)	2.2e-3 (0.0)	1.6e-3 (0.0)	1.4e-3 (0.0)
	Acc. vol (ω_{16})	1.6e-4 (0.01)	2.1e-4 (0.0)	1.4e-4 (0.0)	1.3e-4 (0.67)*
Averaged fitted polynomials of degree 1	Cogn. sc. $y_{\text{fit},1}$	-10.4*t + 226.8	-11.1*t + 320.4	-22.1*t + 243.4	-12.4*t + 187.1
	Mot. sc $y_{\text{fit},2}$	-1.1*t -18.9	-1.9*t + 0.5	-3.6*t -10.2	-5.4*t -18.3
	Psych. sc $y_{\text{fit},3}$	-2.6*t -6.2	0.8*t -16.0	-1.0*t -14.8	0*t -22.4
	Caud. vol $y_{\text{fit},4}$	-1.1e-4*t + 1.6e-3	-3.4e-4*t + 2.6e-3	-2.9e-4*t + 2.0e-3	-2.8e-4*t + 2.0e-3
	Put. vol. $y_{\text{fit},5}$	-2.2e-4*t + 2.2e-3	-4.3e-4*t + 3.3e-3	-3.9e-4*t + 2.7e-3	-3.4e-4*t + 2.5e-3
	Acc. vol. $y_{\text{fit},6}$	-2.1e-5*t + 2.0e-4	-4.0e-5*t + 3.2e-4	-3.8e-5*t + 2.4e-4	-3.2e-5*t + 2.2e-4
	Averaged $y_{\text{fit,all}}$	-2.3*t + 33.6	-2.0*t + 50.8	-4.5*t + 36.4	-3.0*t + 24.4

Table 2. Case 2 averaged feature scores over all patients and times for each cluster (C_l, l = 1–4), for $thres_{gr} = 3.75$ and λ = 0.01. The corresponding distribution of the score weights ω_{kl}, $k = 1 - 6$ is also reported in parenthesis. Subsequently, the averaged fitted polynomials of degree 1 are shown with respect to time for each feature separately averaged over all patients ($y_{\text{fit},k}$, k = 1–6) and averaged over all features and all patients ($y_{\text{fit,all}}$). Higher score values indicate improved functioning. *The asterisk denotes feature scores that contributed more than 5% to the cluster assignment (i.e.$\omega_{kl} > 0.05$); The abbreviated descriptions of the left column refer to the following features (sequentially, see also Table S2): total cognitive score, total motor score, total psychiatric score, caudate volume, putamen volume and nucleus accumbens volume (the three latter denote averaged MRI volumes across right and left hemispheres for each participant).

than in the manifest group, albeit with higher values (i.e., less atrophy) at the time of enrollment ($y_{\text{fit},4}$, $y_{\text{fit},5}$, $y_{\text{fit},6}$), as expected.

By observing the individual Silhouette scores (SS) in Supplementary Table S1, the averaged SS values were overall relatively high and in agreement with the trend observed in the average-distance values provided in the same table (S1) and discussed earlier. In this sense, the most efficient clustering configuration in terms of cohesion and separation was achieved, as expected, for case example 1 (SS > 0.9). SS was reduced for case 4, as it represented the coarsest clustering output among all four.

Discussion

This study presents an innovative data-mining methodological pipeline designed to cluster patients into subgroups that share similar profiles of time-varying multi-modal clinical and imaging features. To demonstrate the potential of the current methodology in elucidating individual heterogeneity hidden beneath a common diagnosis, HD was employed as a model. By analyzing the severity and progression of six target clinical and imaging features in HD patients over time, this work serves as a proof-of-concept in the stratification of patients with a current or impending diagnostic label into groups with similar disease trajectories. The present approach expands on earlier unsupervised-clustering methodology[15,18,19] by incorporating a new dimension of numerical features through an adapted cost-minimization algorithm for clustering on different, possibly overlapping, feature subsets, while also leveraging dynamic time warping to account for variations in the dynamics of disease progression. In this manner, a valuable contribution is provided to the currently limited literature on data-mining techniques for longitudinal cohort analyses in biomedicine.

Methodologically, the study explores various user-defined parameters, such as the granularity threshold and feature contribution parameter (λ), allowing their flexible customization to meet potential specific clinical needs. Specifically, Case 1 was defined by optimum parameters (granularity threshold and feature contribution), as marked by the smallest total average distance, producing ten clusters (Fig. 1). Case 2 shared the same feature contribution parameter as Case 1, but with a higher granularity threshold, in order to produce coarser clustering (i.e., fewer clusters) at the cost of reduced homogeneity. In this case, patient profiles previously assigned to ten clusters (Fig. 1) were merged and/or re-organized to finally form four clusters (Fig. 2). For example, the premanifest cluster C_2 of Case 2 (Fig. 2) included the majority of the premanifest patients from C_8 of Case 1 (Fig. 1), as well as five additional patients (from different clusters of Fig. 1), all exhibiting mild motor symptoms.

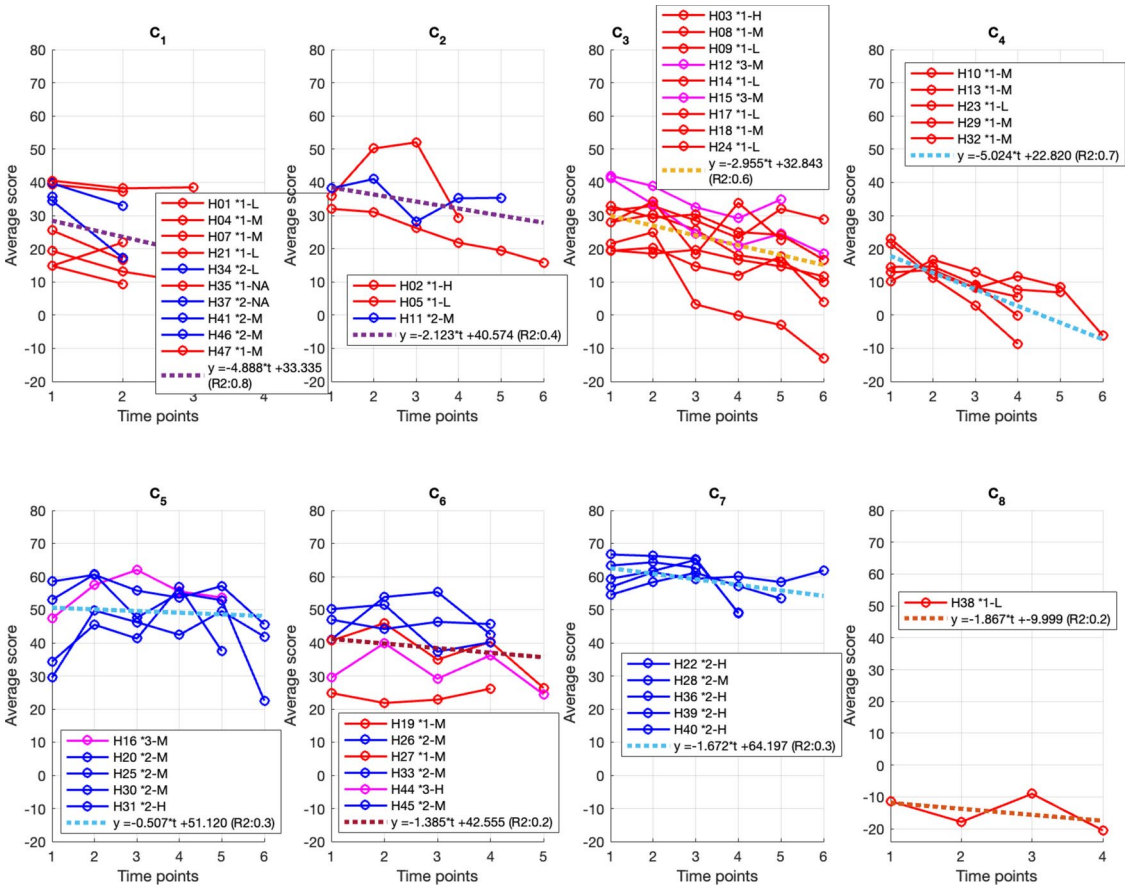

Fig. 3. Extracted clusters of HD trajectories for Case 3 ($thres_{gr} = 3.75$ and $\lambda = 0.4$), versus ordered time points. Red depicts manifest profiles, blue premanifest and pink those individuals that transitioned from premanifest to manifest over the course of assessment. The average score on the y-axis is produced by averaging all six features at each time point. The fitted polynomial of degree 1 is also shown (dotted lines) in each cluster averaged over all patients, together with the R2 (square of the Pearson correlation coefficient).

Case 3 had an identical granularity threshold as in Case 2, with an increased feature contribution parameter, in order to demonstrate the result of permitting a higher number of features to take part in the formation of each cluster. As a result, eight clusters were extracted (Fig. 3) with a resulting decrease in cluster homogeneity, since forming clusters with small dispersion in all or many features is not a practically optimal scenario (see "Methods"). Finally, Case 4 was defined by an increase in the granularity threshold, while maintaining the feature contribution parameter the same as Case 3. This resulted in the coarsest possible separation, dividing patients into a manifest (C_1) and premanifest (C_2) cluster (Fig. 4). In addition, the resulting total average distance was the highest compared to all previous case examples, demonstrating that clusters were the least homogeneous of the case series. Of note, this coarse division between premanifest and manifest is that which is currently utilized in the clinical context, based solely on the motor score and a diagnostic confidence score determined by the clinician. Ultimately, this intra-diagnostic heterogeneity further underscores the potential for the current methodology to further refine diagnostic criteria and prognostic predictions.

While the main focus of the present study was not to provide a detailed clinical interpretation of all extracted patterns or their causality, it establishes a flexible methodological framework aimed at enhancing the understanding of heterogeneity in disease progression. This work serves as a proof-of-concept, demonstrating the potential for personalized patient care by allowing customization of the clustering parameters to meet specific clinical needs. Future studies should replicate and extend this work using other datasets, including large multi-center cohorts such as TRACK-HD or Enroll-HD[37,38], to further explore its applicability. It should be noted that inherent to the nature of the health dataset used, there may be incomplete data and errors in the clinical time registries and/or dates and thus, the results should be interpreted accordingly.

In the post-processing analysis, alternative fitting approaches could be explored to depict the disease's evolution over time more accurately within each cluster, such as polynomials of higher degrees and other non-linear functions. In addition, the current methodology can accommodate the incorporation of static features that may feasibly impact disease progression. Future research may consider investigating the impact of baseline variables on the disease trajectories, such as CAG repeat length[39], cognitive reserve[40], and other covariates such as sex and years of education.

	C_l	1	2	3	4	5	6	7	8
Averaged feature scores	Cogn. sc (ω_{1l})	203.5 (0.421)*	231.2 (0.130)*	177.9 (0.140)*	116.0 (0.297)*	312.8 (0.140)*	251.2 (0.130)*	363.5 (0.248)*	41.8 (0.167)*
	Mot. sc (ω_{2l})	-17.0 (0.173)*	-7.3 (0.227)*	-27.0 (0.165)*	-43.5 (0.215)*	-1.5 (0.298)*	-4.6 (0.230)*	-0.8 (0.326)*	-62.8 (0.167)*
	Psych. sc (ω_{3l})	-18.1 (0.161)*	-20.3 (0.209)*	-14.7 (0.061)*	-15.6 (0.122)*	-17.0 (0.108)*	-14.5 (0.050)	-2.1 (0.270)*	-67.0 (0.167)*
	Caud. vol (ω_{4l})	1.4e-3 (0.090)*	1.7e-3 (0.059)*	1.1e-3 (0.267)*	1.1e-3 (0.178)*	1.6e-3 (0.242)*	1.5e-3 (0.272)*	1.9e-3 (0.046)	1.0e-3 (0.167)*
	Put. vol (ω_{5l})	1.9e-3 (0.084)*	2.1e-3 (0.069)*	1.5e-3 (0.180)*	1.5e-3 (0.125)*	2.0e-3 (0.112)*	2.0e-3 (0.266)*	2.3e-3 (0.026)	1.0e-3 (0.167)*
	Acc. vol (ω_{16})	1.8e-4 (0.071)*	1.78e-4 (0.306)*	1.3e-4 (0.187)*	1.4e-4 (0.063)*	1.9e-4 (0.100)*	1.9e-4 (0.052)*	1.9e-4 (0.084)*	1.0e-4 (0.167)*
Averaged Fitted polynomials of degree 1	Cogn. sc. $y_{\text{fit},1}$	-24.3*t + 228.8	-12.9*t + 270.0	-12.6*t + 221.3	-19.0*t + 166.6	-8.8*t + 343.5	-5.0*t + 266.1	-10.0*t + 387.3	-7.3*t + 60.0
	Mot. sc $y_{\text{fit},2}$	-1.2*t -18.3	-2.3*t -0.7	-4.1*t -13.3	-11.2*t -14.1	-0.6*t + 0.3	-0.7*t -2.6	-0.5*t + 0.6	-4.9*t -50.5
	Psych. sc $y_{\text{fit},3}$	-3.8*t -10.6	2.4*t -25.8	-1.1*t -11.0	0.0*t -15.6	6.4*t -37.1	-2.6*t -8.2	0.5*t -2.8	1.0*t -69.5
	Caud. vol $y_{\text{fit},4}$	-1.0e-4*t + 1.4e-3	-3.8e-4*t + 2.8e-3	-2.4e-4*t + 1.9e-3	-2.8e-4*t + 1.8e-3	-3.6e-4*t + 2.8e-3	-3.4e-4*t + 2.4e-3	-4.2e-4*t + 3.0e-3	-3.2e-4*t + 1.7e-3
	Put. vol. $y_{\text{fit},5}$	-2.3e-4*t + 2.0e-3	-4.8e-4*t + 3.5e-3	-3.0e-4*t + 2.5e-3	-3.3e-4*t + 2.4e-3	-4.5e-4*t + 3.5e-3	-4.3e-4*t + 3.1e-3	-4.9e-4*t + 3.6e-3	-3.4e-4*t + 1.9e-3
	Acc. vol. $y_{\text{fit},6}$	-1.8e-5*t + 1.8e-4	-4.0e-5*t + 3.0e-4	-2.9e-5*t + 2.3e-4	-3.5e-5*t + 2.4e-4	-3.9e-5*t + 3.2e-4	-4.9e-5*t + 3.2e-4	-4.6e-5*t + 3.1e-4	-3.2e-5*t + 1.8e-4
	Averaged $y_{\text{fit,all}}$	-4.9*t + 33.3	-2.1*t + 40.6	-3.0*t + 32.8	-5.0*t + 22.8	-0.5*t + 51.1	-1.4*t + 42.6	-1.7*t + 64.2	-1.9*t -10.0

Table 3. Case 3 averaged feature scores over all patients and times for each cluster (C_l, l = 1–8), for $thres_{gr} = 3.75$ and λ = 0.4. The corresponding distribution of the score weights $\omega_{kl}, k = 1 - 6$ is also reported in parenthesis. Subsequently, the averaged fitted polynomials of degree 1 are shown with respect to time for each feature separately averaged over all patients ($y_{\text{fit},k}$, k = 1–6) and averaged over all features and all patients ($y_{\text{fit,all}}$). Higher score values indicate improved functioning. *The asterisk denotes feature scores that contributed more than 5% to the cluster assignment $(\text{i.e.}\omega_{kl} > 0.05)$; The abbreviated descriptions of the left column refer to the following features (sequentially, see also Table S2): total cognitive score, total motor score, total psychiatric score, caudate volume, putamen volume and nucleus accumbens volume (the three latter denote averaged MRI volumes across right and left hemispheres for each participant).

Considering individual variability constitutes a basic principle of precision medicine, which strives to tailor preventative, diagnostic and therapeutic strategies to the level of a single patient or subgroup of patients. In this manner, future directions could investigate the elaboration of personalized therapeutic management or inclusion of patients with similar features into clinical trials, even in early stages of the disease and across disease categories (e.g., premanifest vs. manifest, pre-diabetic vs. diabetic, Stage I vs. Stage II breast cancer, or depression in bipolar II vs. major depressive disorder).

Conclusions

This research introduces a novel data-mining methodological framework that integrates temporal characteristics of a set of multi-modal clinical and imaging features of patients. The proposed methodology illustrates how temporal differences in the disease progression of patients with the same diagnostic label, in this case with Huntington's disease, can be identified and evaluated. Furthermore, the methodology allows for user-defined flexibility that allows exploration of data through different lenses, while also promoting a quantitative metric, such as the cluster-evaluation process, to assess the optimization of clustering parameters.

The flexible nature of the methodology allows for future expansions that incorporate additional longitudinal patient information, such as prescription drugs or other imaging and laboratory data. Finally, this work can be adapted and applied to other EHRs in the context of RWD, with the objective of further promoting personalized medicine and improving human health.

Methods

Participants

This was a prospective observational study of a cohort of patients undergoing clinical follow-up for Huntington's disease. From 2013 to 2019, 47 participants with Huntington's disease (22 premanifest, 25 manifest) participated in the study and underwent clinical evaluation and MRI. Three participants were excluded due to inability to undergo the MRI scan in the setting of claustrophobia and/or incomplete evaluation for an entire domain, resulting in 44 participants for the cluster analysis. Demographics of the whole sample are detailed in Supplementary Table S2. Premanifest individuals are clinically defined as those who have been

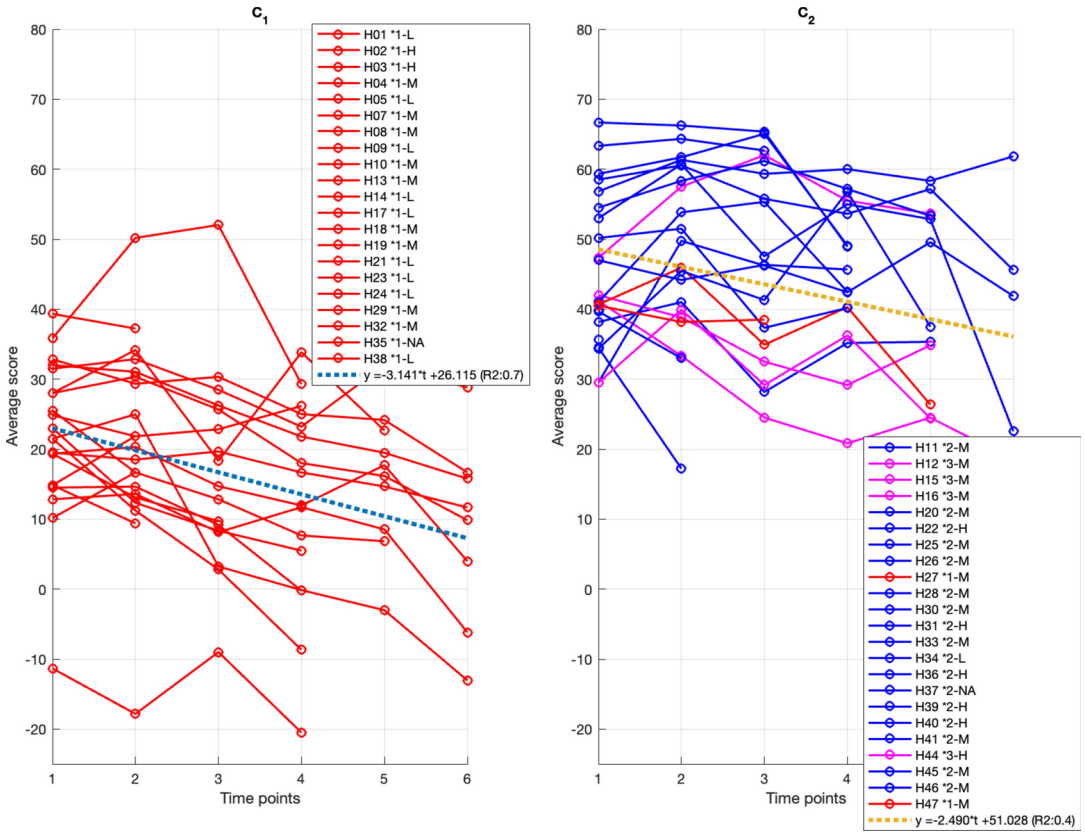

Fig. 4. Extracted clusters of HD trajectories for Case 4 ($thres_{gr} = 6.0$ and $\lambda = 0.4$), versus ordered time points. Red depicts manifest profiles, blue premanifest and pink those individuals that transitioned from premanifest to manifest over the course of assessment. The average score on the y-axis is produced by averaging all six features at each time point. The fitted polynomial of degree 1 is also shown (dotted lines) in each cluster averaged over all patients, together with the R2 (square of the Pearson correlation coefficient).

genetically determined to have the *HTT* gene mutation but have not yet been formally diagnosed with HD by motor criteria. Hence, while not all participants displayed motor symptoms, each was a confirmed HD gene-expansion carrier at baseline (44.02 ± 3.05 CAG repeats). The standardized CAG-Age Product (CAP) score (CAP = 100 × age × (CAG−35.5)/627) was used as a measurement of HD state[39]. All participants were assessed with the Unified Huntington's Disease Rating Scale total motor score (UHDRS-TMS) and total cognitive (UHDRS_cogscore) evaluation[41] (Supplementary Table S2). Neuropsychiatric evaluation was performed with the short-Problem Behavior Assessment (PBA-s)[30]. Details of clinical assessments are provided below. For each participant, there were a maximum total of six longitudinal visits, including baseline (mean number of 4.3 ± 1.6 assessments and mean inter-assessment duration of 14.0 ± 4.1 months). In total, 559 clinical evaluations were obtained. Four premanifest participants converted to manifest over the course of the study.

Clinical evaluation

The UHDRS-cogscore was employed to evaluate phonetic verbal fluency (F-A-S test) and psychomotor speed (Symbol Digit Modalities Test), as well as processing speed, attention and inhibitory control (word-reading, color-naming and interference components of the Stroop Test). To evaluate motor symptomatology, we employed the UHDRS-TMS, which quantifies dysarthria, chorea, dystonia, gait, postural stability and oculomotor function[42].

Neuropsychiatric features were evaluated using the PBA-s[43]. This semi-structured interview is administered in the presence of a knowledgeable informant and consists of eleven components: depressed mood, suicidal ideation, anxiety, irritability, angry or aggressive behavior, lack of initiative (apathy), preservative thinking/behavior, obsessive–compulsive behavior, paranoid thinking/behavior, hallucinations and disoriented thinking/behavior. Domains are calculated as the product of frequency × severity for each sign. Increasing scores in either the motor (UHDRS-TMS) or psychiatric domain (PBA-s) indicate deterioration of the disease. The opposite holds true for the cognitive domain (UHDRS-cogscore), that is, higher score values indicate better performance. In order to align scores and indicate directionality in a consistent way for all three domains, we have multiplied motor and psychiatric scores by -1. In this manner, a more negative score indicates worse evolution. All assessments were carried out by neurologists or neuropsychologists specializing in movement disorders. No participants reported previous history of neurological disorder other than HD.

	C_l	1	2
Averaged feature scores	Cogn. sc (ω_{1l})	155.9 (0.172)	285.5 (0.165)
	Mot. sc (ω_{2l})	-32.1 (0.010)	-3.5 (0.461)
	Psych. sc (ω_{3l})	-18.6 (0.072)	-13.6 (0.080)
	Caud. vol (ω_{4l})	1.1e-3 (0.231)	1.7e-3 (0.104)
	Put. vol (ω_{5l})	1.5e-3 (0.228)	2.1e-3 (0.088)
	Acc. vol (ω_{16})	1.4e-4 (0.197)	2.0e-4 (0.103)
Averaged fitted polynomials of degree 1	Cogn. sc. $y_{\text{fit},1}$	-13.4*t + 192.3	-13.9*t + 320.8
	Mot. sc $y_{\text{fit},2}$	-4.9*t-18.3	-1.0*t -0.8
	Psych. sc $y_{\text{fit},3}$	-0.6*t -17.4	-0.1*t -13.9
	Caud. vol $y_{\text{fit},4}$	-2.0e-4*t + 1.7e-3	-3.5e-4*t + 2.6e3
	Put. vol. $y_{\text{fit},5}$	-2.9e-4*t + 2.3e-3	-4.4e-4*t + 3.3e3
	Acc. vol. $y_{\text{fit},6}$	-2.8e-5*t + 2.1e-4	-4.1e-5*t + 3.0e4
	Averaged $y_{\text{fit,all}}$	-3.1*t + 26.1	-2.5*t + 51.0

Table 4. Case 4 averaged feature scores over all patients and times for each cluster (C_l, l = 1,2), for $thres_{gr} = 6.0$ and λ = 0.4. The corresponding distribution of the score weights ω_{kl}, $k = 1 - 6$ is also reported in parenthesis. Subsequently, the averaged fitted polynomials of degree 1 are shown with respect to time for each feature separately averaged over all patients ($y_{\text{fit},k}$, k = 1–6) and averaged over all features and all patients ($y_{\text{fit,all}}$). Higher score values indicate improved functioning. * The asterisk denotes feature scores that contributed more than 5% to the cluster assignment (i.e.$\omega_{kl} > 0.05$); The abbreviated descriptions of the left column refer to the following features (sequentially, see also Table S2): total cognitive score, total motor score, total psychiatric score, caudate volume, putamen volume and nucleus accumbens volume (the three latter denote averaged MRI volumes across right and left hemispheres for each participant).

MRI acquisition and processing

MRI data were acquired with a 3 T whole-body MRI scanner (Siemens Magnetom Trio; Hospital Clinic, Barcelona), using a 32-channel phased array head coil to procure structural T1-weighted images (magnetization-prepared rapid-acquisition gradient echo sequence), 208 sagittal slices, repetition time = 1970 ms, echo time = 2.34 ms, inversion time = 1050 ms, flip angle = 9º, field of view = 256 mm, 1 mm isotropic voxel with no gap between slices.

Subcortical volumes for three brain structures, namely the caudate nucleus, putamen and nucleus accumbens, were estimated for each participant through an automated procedure for volumetric segmentation of the T1-weighted images using FreeSurfer (https://surfer.nmr.mgh.harvard.edu/). This parcellation allows the extraction of subcortical volumes that has previously been associated with neurodegenerative processes, including the total intracranial volume (TIV). TIV is used as a normalization factor, adjusting for head size differences. Volumes were averaged across the left and right for each anatomical region. For each participant, MRI data from a maximum of two longitudinal visits was obtained, over a period of 18.1 ± 6.1 months.

Formation of HD trajectories

For an individual under consideration, the HD trajectory was defined as a finite time sequence of T chronologically ordered registries of clinical or MRI data as described above. A list of K = 6 clinical assessment scores was assigned at each time registry (Supplementary Table S3). These include the clinical evaluations (UHDRS-cogscore, UHDRS-TMS and PBA-s), as well as the three MRI-based volumes (caudate, putamen and nucleus accumbens), the latter of which are normalized by the TIV and then averaged over the corresponding left and right hemispheres. The aforementioned scores will be denoted hereafter as *features*, which form a *feature vector* at each time instant.

Subsequently, the HD trajectories corresponding to patients $i = 1,2,\ldots,n$, can be written as:

$$p^{(i)} = \left[p_{jk}^{(i)}\right], \quad \text{for } k = 1,2,\ldots,K \quad \text{and } j = 1,2,\ldots,T. \tag{1}$$

where the elements of the above temporal sequence represent registries at discrete times $t_j^{(i)}$ and k denotes the features assigned each time and $T \geq 1$. Equation (1) can be expanded in two dimensions, yielding the following $K \times T$ matrix for each HD patient trajectory:

$$p^{(i)} = \begin{bmatrix} p_{11}^{(i)} & p_{21}^{(i)} & \cdots & p_{N1}^{(i)} \\ p_{12}^{(i)} & p_{22}^{(i)} & & p_{N2}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1K}^{(i)} & p_{2K}^{(i)} & \cdots & p_{NK}^{(i)} \end{bmatrix}, \quad \text{for } i = 1, 2, \ldots, n. \tag{2}$$

where $n = 44$.

Due to the variability in the number of longitudinal registries between all features, the resampling method was applied on each feature and patient (method *resample* in MATLAB) in order to equalize the time length of all six feature vectors. This was achieved by upsampling each feature vector at L_p/L_o times the original sample rate, where L_p denoted the maximum longitudinal length among all six feature vectors corresponding to a specific patient and L_o the length of the original feature.

DTW-based clustering of HD trajectories

Subsequently, the clinical trajectories of all HD patients were clustered in order to identify subgroups of patient profiles with shared temporal patterns. For this objective, the disease-trajectory clustering algorithm presented in[15,18] was adapted in order to incorporate not only the time, but also the feature dimension described earlier. In order to incorporate the time-varying feature vector into the clustering algorithm, the previous technique[20] for clustering objects on possibly overlapping subsets of features was modified. A schematic illustrating the fundamentals of the proposed methodology is shown in Fig. 5.

In the adapted DTW-based clustering method proposed in this work, an HD trajectory was iteratively compared with the trajectories of the previously formed clusters by calculating their in-between distance. For two patients *i* and *q* with trajectories $p^{(i)}$ and $p^{(q)}$, respectively (described by Eqs. (1) and (2)), the total distance between the two trajectories can be written as follows:

$$D_{iq}\left(t_j^{(i)}, t_{j'}^{(q)}\right) = \sum_{k=1}^{K} \omega_k d_{iqk}\left(t_j^{(i)}, t_{j'}^{(q)}\right), \text{ with } \{\omega_k \geq 0\}_1^K \text{ and } \sum_{k=1}^{K} \omega_k = 1. \tag{3}$$

Time instants $t_j^{(i)}$ and $t_{j'}^{(q)}$ can be distinct in two patients, since the DTW algorithm permits variability in time scaling, length and in-between intervals, as noted in Introduction. The distances d_{iqk}, of which the above equation is composed, refer to the different features under investigation and are given by:

$$d_{iqk}\left(t_j^{(i)}, t_{j'}^{(q)}\right) = \frac{\left|p_{jk}^{(i)} - p_{j'k}^{(q)}\right|}{s_k}, \text{ for } k = 1, \ldots, K. \tag{4}$$

The denominator s_k provides a scale for measuring "closeness" (spread) on each feature *k* over all patients of the cohort, i.e.:

$$s_k = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{q=1}^{n} \left| p_{\bar{j}k}^{(i)} - p_{\bar{j}'k}^{(q)} \right| \tag{5}$$

where the absolute distance is calculated, in a pair-wise manner for all patients of the cohort, at averaged times $t_{\bar{j}}^{(i)}, t_{\bar{j}'}^{(q)}$ for each pair of patients. As observed in (4), the issue of varying feature magnitudes is addressed in the algorithm by dividing the individual distances by dispersion s_k, which ensures the establishment of a comparable range for all features across the clustering process.

The relative influence of each score $p_{jk}^{(i)}$ is regulated by its corresponding weight ω_k in (3). Feature selection seeks to find an optimal weighting $\boldsymbol{\omega} = \{\omega_k \geq 0\}_1^K$ as part of the clustering problem by jointly minimizing the clustering criteria set in[20]. Although feature selection can be helpful, it attempts to form clusters based on the same subset of features. However, there may exist clusters whose groupings are formed based on different and possibly overlapping feature subsets. In this work, this kind of clustering was sought for the flexibility that it provides. For this reason, the generalized formula resulting from (3) was used, such that a separate score weighting $\boldsymbol{\omega}_l = \{\omega_{kl} \geq 0\}_1^K$ was defined for each cluster C_l. As previously, the score weights satisfy $\{\omega_{kl} \geq 0\}_1^K$ and $\sum_{k=1}^{K} \omega_{kl} = 1$.

Finally, since the objective of this work was to extract clusters of patients that cluster simultaneously on subsets of scores (features), where each subset contains more than one score, the distance between two HD sequences $p^{(i)}$ and $p^{(q)}$, becomes:

$$D_{iq}\left(t_j^{(i)}, t_{j'}^{(q)}\right)_{\boldsymbol{\omega}_l} = \sum_{k=1}^{K} \left(\omega_{kl} d_{iqk}\left(t_j^{(i)}, t_{j'}^{(q)}\right) + \lambda \omega_{kl} log \omega_{kl}\right) + \lambda log K \tag{6}$$

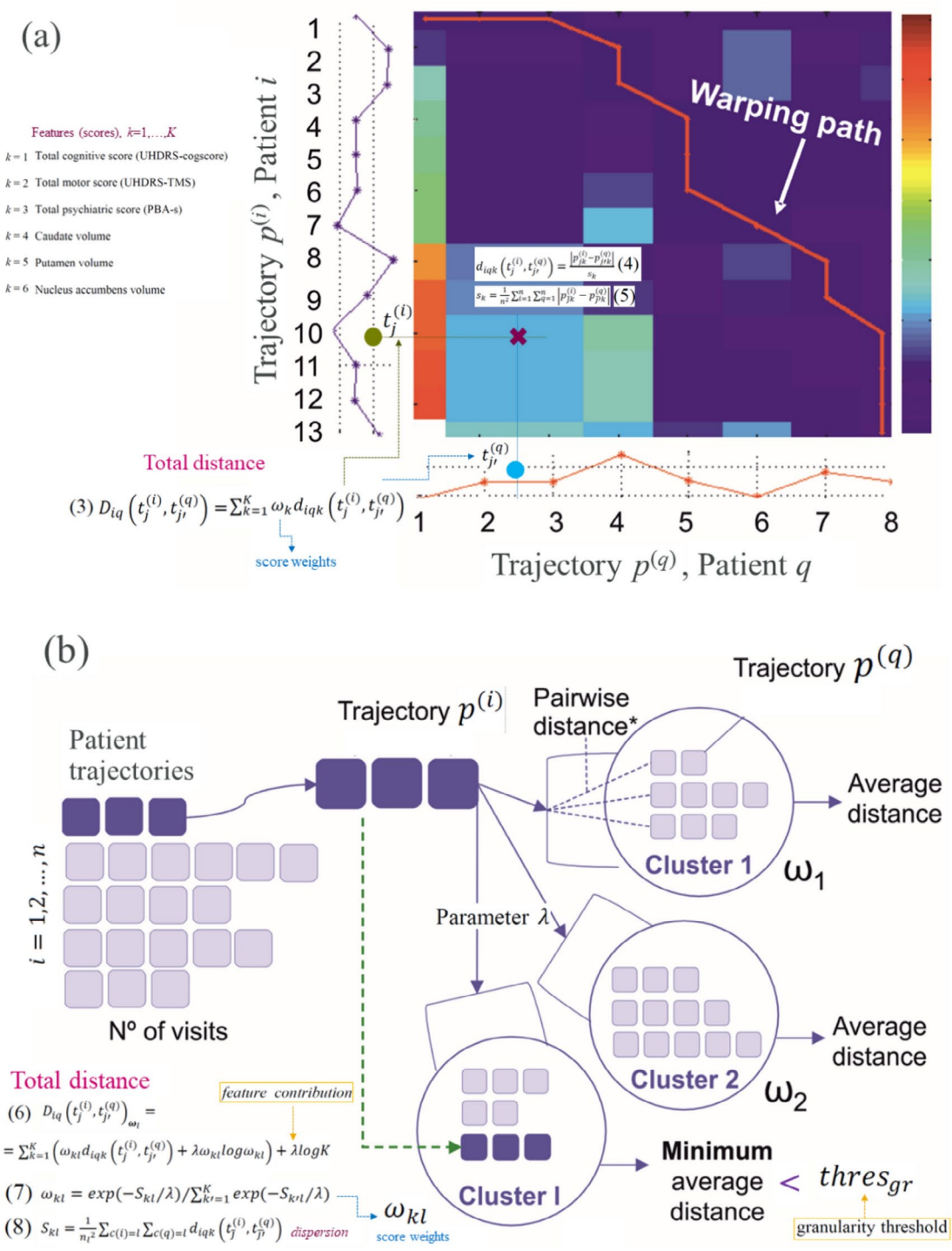

Fig. 5. A schematic illustrating the fundamental aspects of the proposed DTW-based clustering methodology: (**a**) The calculation of the total distance $D_{iq}\left(t_j^{(i)}, t_{j'}^{(q)}\right)$ (Eq. (3)) between two HD trajectories $p^{(i)}$ and $p^{(q)}$, defined as sequences of K time-varying features (scores, left panel), is visually described through the construction of the local distance matrix and derivation of the warping path [15]. For each coordinate point in the above matrix, the absolute distance between the trajectories at time points $t_j^{(i)}, t_{j'}^{(q)}$ and feature k is considered (Eq. (4)), normalized by s_k (Eq. (5)). The total distance is finally given as a weighted sum, with weights ω_k reflecting the relative influence of each score. The colorbar in the figure denotes distance ranging from lowest (dark blue) to highest (dark red). (**b**) The total distance is next plugged into the unsupervised clustering algorithm, a snapshot of which is shown at a random iteration. Trajectory $p^{(i)}$ is assigned to the cluster with the minimum average distance between this and all other trajectories $p^{(q)}$ already assigned to other clusters. Furthermore, by tuning the feature contribution parameter λ and granularity threshold $thres_{gr}$, different clustering configurations can be achieved depending on the requirements of a study

where the score weights are described by:

$$\omega_{kl} = exp\left(-S_{kl}/\lambda\right) / \sum_{k\prime=1}^{K} exp\left(-S_{k\prime l}/\lambda\right) \tag{7}$$

with $S_{kl}, S_{k\prime l}$ given by

$$S_{kl} = \frac{1}{{n_l}^2} \sum_{c(i)=l} \sum_{c(q)=l} d_{iqk}\left(t_{\bar{j}}^{(i)}, t_{\bar{j}\prime}^{(q)}\right) \tag{8}$$

S_{kl} is a measure of the dispersion (scale) of the data values on the k-th score for patients in the l-th cluster and n_l defines the corresponding number of patients assigned to it.

At each iteration of the clustering algorithm, the score weights ω_{kl} were calculated for each cluster and then the minimum average distance between each new HD trajectory and the trajectories already assigned to each cluster was sought. The algorithm terminated when there were no other HD trajectories to be considered.

User-defined parameters and cluster evaluation

The DTW-based clustering algorithm presented above involved the selection of two user-defined parameters. First, the parameter λ seen in (6)-(7) controlled for the incentive for clustering on more features ($\lambda \geq 0$), such that a lower or higher value of λ encouraged clusters on less or more scores, respectively. The score weights described by (7) put higher weight on features k with smaller dispersion (S_{kl}) within each cluster l. Setting $\lambda = 0$ puts all weight on the feature with the smallest S_{kl}, while $\lambda = \infty$ forces all features to contribute with equal weight to each cluster. Parameter λ will be denoted hereafter as *feature contribution parameter*. According to [20], there is no known mathematical formula to provide an optimal λ and therefore its value is typically selected based on empirical evaluation. In our clustering simulations, several values of λ were tested in order to assess different clustering configurations according to the degree of contribution for each score.

Furthermore, a granularity threshold $thres_{gr}$ was employed with the purpose of adjusting the clustering granularity. In this sense, lower values of $thres_{gr}$ increase cluster homogeneity at the cost of increasing the total number of clusters and vice versa. A compromise was generally sought to achieve sufficiently homogeneous clusters while avoiding excessive cluster fragmentation[15].

In order to optimize the selection of the above user-defined parameters, a cluster evaluation process was proposed that attempts to obtain a quantitative metric of the cluster homogeneity and, consequently, to assess the overall efficiency of the clustering algorithm. This involved the evaluation of different combinations of the granularity threshold and feature contribution parameter (λ) to identify clusters within HD trajectories. Subsequently, the total average distance over all extracted clusters was calculated, with lower values indicating increased homogeneity. In brief, for each combination of $(thres_{gr}, \lambda)$, the DTW-based unsupervised algorithm was applied and the cluster-based average distance was estimated for each identified cluster. This was achieved by averaging the pair-wise distances between all HD trajectories assigned to the cluster under consideration based on the formula described in (6). The score weights ω_{kl} used in that formula are the ones formed at the final step of the iterative clustering algorithm. The total average distance, corresponding to the examined pair of values $(thres_{gr}, \lambda)$, is then obtained by averaging across all extracted clusters.

Finally, the combination(s) of $thres_{gr}$ and λ that minimized the total average distance (thereby increasing cluster homogeneity), while controlling for the total number of extracted clusters was sought. In light of this, a compromise between the aforementioned conditions was needed in order to select the optimal user-defined parameters for the clustering algorithm. These parameters were empirically set to require (i) the total average distance to be < 30% of the maximum distance (resulting from all tested parameter values) and (ii) limiting the number of outlier clusters (those containing a single trajectory) to < 25% of the maximum number of extracted clusters.

Furthermore, an adapted Silhoutte score (SS) was employed to provide an additional measure of how well-defined the clusters are[44]. It was calculated based on the formula: $(b\text{-}a)/\max(a,b)$, where a denotes the average (pair-wise) distance from a trajectory to other trajectories within the same cluster, representing the *cohesion* of the trajectory with its own cluster. b is the average distance from a trajectory to all trajectories in any other cluster, representing the *separation* of the trajectory from other clusters. For the derivation of the SS, the distance in (6) was employed as described above. The range of SS is from -1 to 1, with values closer to 1 indicating well-defined clusters (increased cohesion and separation).

In addition to the previously described evaluation metrics, alternative metrics could be also considered that may align directly with clinical relevance. In this context, features that assess the association between the identified clusters and disease severity, motor symptoms, functional decline, or cognitive impairment could be explored. Specific metrics for analysis may focus on overall severity (i.e., amplitude) of scores at baseline or across all visits, the rate of progression over time (e.g., linear or nonlinear modelling), or within-group dispersion.

Linear fitting of the clustered HD trajectories

In the final step of the proposed methodological pipeline, an approximated linear characterization of the temporal patterns for the identified clusters was performed. Linear progression has been previously modeled in HD for volume loss in the caudate and putamen[45] as well as motor and functional scores of the UHDRS[46,47]. This step provided an approximation of the overall (averaged) progression of the disease as reflected in each cluster and thus facilitated interpretability and comparisons between them. Specifically, linear fitting was achieved by

fitting each HD trajectory assigned to a cluster to a polynomial of degree 1 and averaging these over all the included trajectories. Consequently, seven approximated linear functions of time were obtained for each cluster, corresponding to (i) each feature trajectory individually ($y_{\text{fit},k}$, $k=1$–6) and (ii) an averaged trajectory over all features ($y_{\text{fit,all}}$).

Data availability

The raw data supporting the findings of this study cannot be shared publicly as they contain clinical and genetic data sensitive to the Institution. In the interest of minimizing the risk of participant identification, we will make the data available upon reasonable request to the corresponding author (email: alexia.giannoula@astrazeneca.com), with approval by the local institutional review board.

Received: 12 September 2024; Accepted: 13 January 2025
Published online: 24 January 2025

References

1. Cirillo, D. & Valencia, A. Big data analytics for personalized medicine. *Curr. Opin. Biotechnol.* **58**, 161–167. https://doi.org/10.1016/j.copbio.2019.03.004 (2019).
2. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522. https://doi.org/10.1038/nrg.2016.86 (2016).
3. Hodson, R. Precision medicine. *Nature* **537**, S49–S49. https://doi.org/10.1038/537S49a (2016).
4. Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. Big data in healthcare: management, analysis and future prospects. *J. Big Data* **6**, 54. https://doi.org/10.1186/s40537-019-0217-0 (2019).
5. Ahmad, P., Qamar, S. & Qasim Afser Rizvi, S. Techniques of data mining in healthcare: A review. *IJCA* **120**, 38–50. https://doi.org/10.5120/21307-4126 (2015).
6. Wu, W.-T. et al. Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Mil. Med. Res.* **8**, 44. https://doi.org/10.1186/s40779-021-00338-z (2021).
7. Chen, J. H., Podchiyska, T. & Altman, R. B. OrderRex: Clinical order decision support and outcome predictions by data-mining electronic medical records. *J. Am. Med. Inform. Assoc.* **23**, 339–348. https://doi.org/10.1093/jamia/ocv091 (2016).
8. Hidalgo, C. A., Blumm, N., Barabási, A.-L. & Christakis, N. A. A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1000353 (2009).
9. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 1–10. https://doi.org/10.1038/srep26094 (2016).
10. Chicco, D. & Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med. Inform. Decis. Mak.* **20**, 1–16. https://doi.org/10.1186/s12911-020-1023-5 (2020).
11. Kaur, I., Doja, M. N. & Ahmad, T. Data mining and machine learning in cancer survival research: An overview and future recommendations. *J. Biomed. Inform.* **128**, 104026. https://doi.org/10.1016/j.jbi.2022.104026 (2022).
12. Campbell, E. A., Bass, E. J. & Masino, A. J. Temporal condition pattern mining in large, sparse electronic health record data: A case study in characterizing pediatric asthma. *J. Am. Med. Inform. Assoc.* **27**, 558–566. https://doi.org/10.1093/jamia/ocaa005 (2020).
13. Nagamine, T. et al. Data-driven identification of heart failure disease states and progression pathways using electronic health records. *Sci. Rep.* **12**, 1–20. https://doi.org/10.1038/s41598-022-22398-4 (2022).
14. Jensen, A. B. et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.* **5**, 1–10. https://doi.org/10.1038/ncomms5022 (2014).
15. Giannoula, A., Gutierrez-Sacristán, A., Bravo, Á., Sanz, F. & Furlong, L. I. Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study. *Sci. Rep.* **8**, 1–14. https://doi.org/10.1038/s41598-018-22578-1 (2018).
16. Haug, N. et al. Decompression of multimorbidity along the disease trajectories of diabetes mellitus patients. *Front. Physiol.* https://doi.org/10.3389/fphys.2020.612604 (2020).
17. Vetrano, D. L. et al. Twelve-year clinical trajectories of multimorbidity in a population of older adults. *Nat. Commun.* **11**, 1–9. https://doi.org/10.1038/s41467-020-16780-x (2020).
18. Giannoula, A., Centeno, E., Mayer, M.-A., Sanz, F. & Furlong, L. I. A system-level analysis of patient disease trajectories based on clinical, phenotypic and molecular similarities. *Bioinformatics* **37**, 1435–1443. https://doi.org/10.1093/bioinformatics/btaa964 (2021).
19. Giannoula, A. et al. Exploring long-term breast cancer survivors' care trajectories using dynamic time warping-based unsupervised clustering. *J. Am. Med. Inform. Assoc.* https://doi.org/10.1093/jamia/ocad251 (2024).
20. Friedman, J. H. & Meulman, J. J. Clustering objects on subsets of attributes (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66**, 815–849. https://doi.org/10.1111/j.1467-9868.2004.02059.x (2004).
21. Waldvogel, H. J., Kim, E. H., Thu, D. C. V., Tippett, L. J. & Faull, R. L. M. New perspectives on the neuropathology in Huntington's disease in the human brain and its relation to symptom variation. *J. Huntington's Dis.* **1**, 143–153. https://doi.org/10.3233/JHD-2012-120018 (2012).
22. Tabrizi, S. J. et al. Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. *Lancet Neurol.* **10**, 31–42. https://doi.org/10.1016/S1474-4422(10)70276-3 (2011).
23. Waldvogel, H. J., Kim, E. H., Tippett, L. J., Vonsattel, J.-P.G. & Faull, R. L. The neuropathology of Huntington's disease. In *Behavioral Neurobiology of Huntington's Disease and Parkinson's Disease* (eds Nguyen, H. H. P. & Cenci, M. A.) 33–80 (Springer Berlin Heidelberg, 2014). https://doi.org/10.1007/7854_2014_354.
24. MacDonald, M. E. et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983. https://doi.org/10.1016/0092-8674(93)90585-E (1993).
25. Langbehn, D. R., Hayden, M. & Paulsen, J. S. CAG-repeat length and the age of onset in Huntington disease (HD): A review and validation study of statistical approaches. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B**, 397–408. https://doi.org/10.1002/ajmg.b.30992 (2010).
26. Blumenstock, S. & Dudanova, I. Cortical and striatal circuits in Huntington's disease. *Front. Neurosci.* **14**. https://www.frontiersin.org/article/10.3389/fnins.2020.00082 (2020) (accessed January 15, 2022).
27. Bonelli, R. M. & Cummings, J. L. Frontal-subcortical circuitry and behavior. *Dialogues Clin. Neurosci.* **9**, 141 (2007).
28. Ca, R. & Sj, T. Huntington's disease: from molecular pathogenesis to clinical treatment. *Lancet Neurol.* https://doi.org/10.1016/S1474-4422(10)70245-3 (2011).
29. Cheng, J., Liu, H.-P., Lin, W.-Y. & Tsai, F.-J. Identification of contributing genes of Huntington's disease by machine learning. *BMC Med. Genomics* **13**, 176. https://doi.org/10.1186/s12920-020-00822-w (2020).
30. Diehl, S. K. et al. Motor speech patterns in Huntington disease. *Neurology* **93**, e2042–e2052. https://doi.org/10.1212/WNL.0000000000008541 (2019).

31. Wang, D., Ma, X., Schulz, P. E., Jiang, X. & Kim, Y. Clinical outcome-guided deep temporal clustering for disease progression subtyping. *J. Biomed. Inform.* **158**, 104732. https://doi.org/10.1016/j.jbi.2024.104732 (2024).
32. Ramamoorthy, D. et al. Identifying patterns in amyotrophic lateral sclerosis progression from sparse longitudinal data. *Nat. Comput. Sci.* **2**, 605–616. https://doi.org/10.1038/s43588-022-00299-w (2022).
33. Ko, J. et al. Clustering and prediction of disease progression trajectories in Huntington's disease: An analysis of Enroll-HD data using a machine learning approach. *Front. Neurol.* **13**, 1034269. https://doi.org/10.3389/fneur.2022.1034269 (2023).
34. Raschka, T. et al. Unraveling progression subtypes in people with Huntington's disease. *EPMA J.* **15**, 275–287. https://doi.org/10.1007/s13167-024-00368-2 (2024).
35. Tan, H. H. G. et al. MRI clustering reveals three ALS subtypes with unique neurodegeneration patterns. *Ann. Neurol.* **92**, 1030–1045. https://doi.org/10.1002/ana.26488 (2022).
36. Poulakis, K. et al. Japanese Alzheimer's Disease Neuroimaging Initiative, Australian Imaging, Biomarkers and Lifestyle study, Multi-cohort and longitudinal Bayesian clustering study of stage and subtype in Alzheimer's disease. *Nat. Commun.* **13**, 4566. https://doi.org/10.1038/s41467-022-32202-6 (2022).
37. Landwehrmeyer, G. B. et al. Data analytics from Enroll-HD, a global clinical research platform for Huntington's disease. *Mov. Disord. Clin. Pract.* **4**, 212–224. https://doi.org/10.1002/mdc3.12388 (2016).
38. Tabrizi, S. J. et al. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet Neurol.* **12**, 637–649. https://doi.org/10.1016/S1474-4422(13)70088-7 (2013).
39. Ross, C. A. et al. Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nat. Rev. Neurol.* **10**, 204–216. https://doi.org/10.1038/nrneurol.2014.24 (2014).
40. De Paepe, A. E. et al. Cognitive engagement may slow clinical progression and brain atrophy in Huntington's disease. *Sci. Rep.* **14**, 30156. https://doi.org/10.1038/s41598-024-76680-8 (2024).
41. Unified Huntington's disease rating scale: Reliability and consistency. *Movement Disorders* **11**, 136–142. https://doi.org/10.1002/mds.870110204 (1996).
42. Reilmann, R., Schubert, R. Motor outcome measures in Huntington disease clinical trials. In *Handbook of Clinical Neurology* 209–225 (Elsevier, 2017) https://doi.org/10.1016/B978-0-12-801893-4.00018-3.
43. McNally, G., Rickards, H., Horton, M. & Craufurd, D. Exploring the validity of the short version of the Problem Behaviours Assessment (PBA-s) for Huntington's disease: A Rasch analysis. *J. Huntington's Dis.* **4**, 347–369 (2015).
44. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7 (1987).
45. Tabrizi, S. J. et al. A biological classification of Huntington's disease: the Integrated Staging System. *Lancet Neurol.* **21**, 632–644. https://doi.org/10.1016/S1474-4422(22)00120-X (2022).
46. Tang, C. & Feigin, A. Monitoring Huntington's disease progression through preclinical and early stages. *Neurodegener. Dis. Manage.* **2**, 421. https://doi.org/10.2217/nmt.12.34 (2012).
47. Meyer, C. et al. Rate of change in early Huntington's disease: A clinicometric analysis. *Mov. Disord.* **27**, 118–124. https://doi.org/10.1002/mds.23847 (2012).

Acknowledgements

The authors are grateful to the patients and their families for their participation in this project. We would also like to thank Dr. Saül Martinez-Horta, Dra. Andrea Horta-Barba Dr. Jesús Pérez Pérez, Dr. Jaime Kulisevksy, Pilar Sanchez, Dr. Esteban Muñoz, Celia Mareca, Dr. Ruiz-Idiago, Dra. Matilde Calopa, Nadia Rodriguez-Dechichá, Irene Vaquer, Yemila Plana, Dra. Matilde Calopa and Dra. Clara Garcia-Gorro for help with clinical evaluation and data collection.

Author contributions

A. G. designed the methodology, performed the simulations and analysis of data and drafted the manuscript. A. E. D. helped conceive the project idea and drafted parts of the paper relevant to Huntington's disease, the study participants and their clinical/MRI evaluation. She also revised the paper and helped in interpreting the results. L. F. and F. S. revised and approved the manuscript. E. C. revised the manuscript, helped in interpreting the results and supervised the work.

Funding

IMPaCT-Data (IMP/00019) funded by the Institute of Health Carlos III, co-funded by the European Union, European Regional Development Fund (ERDF, "A way to make Europe"). A. E. D. received funding from the Masters in Multidisciplinary Research in Experimental Sciences of the Barcelona Institute of Science and Technology and University of Pompeu Fabra. E. C. was supported by the Instituto de Salud Carlos III, an agency of the Ministerio de Ciencia, Innovacion y Universidades (MINECO), co-funded by FEDER funds/European Regional Development Fund (ERDF) – a Way to Build Europe (CP13/00225 and PI14/ 00834, to EC), as well as Ministerio de Ciencia e Innovación, which is part of Agencia Estatal de Investigación (AEI), through the Retos Investigación grant, number PID2020-114518RB-I00 / DOI: https://doi.org/10.13039/501100011033 to EC, BFU2017-87109-P, to Ruth de Diego. We thank CERCA Programme/Generalitat de Catalunya for institutional support.

Declarations

Competing interests

A. G. declares being an employee of AstraZeneca LLC and may hold stock.

Ethics Approval

All methods were carried out in accordance with the Helsinki Declaration of 1975. The study was approved by the ethics committee of Bellvitge Hospital. All participants provided written informed consent to participate in the study.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86686-5.

Correspondence and requests for materials should be addressed to A.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.