Behavioral Clusters and Lesion Distributions in Ischemic Stroke, based on NIHSS Similarity Network

Louis Fabrice Tshimanga[1,2*†], Andrea Zanola[1,2*†], Silvia Facchini[1], Antonio Luigi Bisogno[1], Lorenzo Pini[2,4], Manfredo Atzori[1,2,3], Maurizio Corbetta[1,2,4]

[1*]Department of Neuroscience, University of Padua, Padua, 35128, Italy.
[2*]Padova Neuroscience Center, University of Padua, Padua, 35128, Italy.
[3*]Information Systems Institute, University of Applied Sciences Western Switzerland, Sierre, 3960, Switzerland.
[4*]Veneto Institute of Molecular Medicine, Padua, 35128, Italy.

*Corresponding author(s). E-mail(s): louisfabrice.tshimanga@unipd.it;
†These authors contributed equally to this work.

Abstract

Purpose: Stroke is a leading cause of death and disability. The subsequent dysfunctions relate to brain lesion location. The complex relationship between behavioral symptoms and lesion location is essential for care and rehabilitation, and useful to understand the healthy brain. Such complexity eludes linear methods. Differently from previous works, this study clusters patients by recurrent symptoms profiles, then retrieves their neural correlates.

Methods: Prevalence, severity and associations of symptoms affecting patients are condensed in a similarity measure between their profiles. The computation specifically adheres to the ordinal nature of NIHSS data. A network model links patients with strength proportional to the behavioral similarity. The network nodes are then classified through a novel spectral clustering variation. Lesions from each cluster's patients are used in a voxel-wise analysis to statically validate the findings.

Results: The behavioral clusters express symptoms co-occurring in accordance with the literature on behavioral deficits covariance. Moreover, they show coherent groups of brain lesions in the associated CT and MRI scans. The anatomical position of the significant voxels found are aligned with the literature. Even when lesion density maps overlap, the significant voxels are well separated.
Conclusion: The presented workflow offers concrete, clinically useful correlations between patients profiles, individual patients-to-group associations, in-group co-occurrences of symptoms, and brain lesion correlates. The underlying mathematical assumptions, and the unsupervised machine learning techniques offer statistically robust results and are being further applied to other multimodal biomedical data.

Keywords: Stroke, NIHSS, Machine Learning, Clustering

1 Introduction

Stroke is the second leading cause of death and the third leading cause of disability worldwide [1] [2]. It occurs when there is an interruption in oxygenation to the brain, leading to cell death. In ischemic stroke, blood clots interrupt blood flow, whereas in hemorrhagic stroke, vessels break and blood diverges from its intended path. The subsequent damage to the brain of survivors can result in acute and chronic dysfunctions across sensory, motor, cognitive and behavioural domains. For instance, focal cortical lesions in highly specialized brain regions can affect narrow sets of functions, as in Broca's and Wernicke's aphasias. More often, subcortical strokes damage both grey and white matter, disrupting connections across the brain beyond the lesion location. These phenomena shed light and spark debates on the functional organization of the brain.

To characterize and quantify behavioral symptoms, clinicians have developed several tests that score performance in multiple domains, such as the Montreal Cognitive Assessment (MoCA) [3], the Oxford Cognitive Screen (OCS) [4] and the National Institutes of Health Stroke Scale (NIHSS) [5] [6]. The NIHSS is a 15 item scale that rates the severity of sensory, motor, attention and language deficits in stroke survivors. A

score of 0 signifies the absence of deficit and healthy response, while maximum damage values range from 2 to 4, depending on the item. For example, the motor arm test involves evaluating the ability to hold the limb against gravity; 0 is given when there is no drift (limb hold for full 10 seconds at 45 or 90 degrees), while 4 is given when the patient is unable to move the limb; each limb has its respective NIHSS item. The NIHSS test can be quickly administered at admission, during the acute phase, at the neuropsychological evaluation, and later on during recovery. Many studies have focused on the sum of all item scores, while analyses of the separate scores have often characterized only inter-item correlations at the population level.

Such analyses have repeatedly identified a low-dimensional structure of the behavioral deficits' covariance, with 2 up to 5 [7] [8] [9] [10] [11] components explaining the larger share of variance. These factors of variability separately correlate with lateral motor impairments, language, memory, and attention. Regularized regression models, statistical tests and machine learning techniques have been developed to relate behavioural scores to brain scans.

Despite several studies have analysed the relationship between lesion locations and behavioural symptoms, some delimiting aspects should be highlighted. First, many studies rely on implicit mathematical assumptions that might distort or discard information. In particular, although the use of rank correlations and tests is settled [12] [13], ordinal scores as found in NIHSS are usually treated as counts or continuous intervals [14] [15]. Alternatively, ordinal scales are reduced to dichotomous variables, e.g. $\{0, 1\}$ to denote absence or presence of deficits [16] [17], or the cumulative sum of scores is thresholded to separate into classes of milder impairment and classes of worse prognosis [15] [18].

Second, the relationships among item scores [7] [8] and with lesion volumes and locations are usually modeled with linear methods, both for ordinal [19] and continuous scores [20].

Third, the scope is usually to study global, population-level associations between single deficits, rather than conditional associations and multivariate distributions of symptoms, i.e. syndromes. The former perspective, while insightful especially when developing clinical tests, is not concerned with individual-level phenotypization, nor sufficient for prediction of groups of deficits. In this regard, it is also notable that predicting deficits from lesions aligns with the aetiological process, but reverses the chronological order of assessment and availability typical of the clinical setting.

This study complements previous perspectives by introducing several innovative approaches. First, a network model of pairwise similarities among subjects is presented, with a conservative distance measure with regards to the ordinal nature of data, better suited than traditional alternatives such as Euclidean or Manhattan/Hamming distances between score vectors. Second, cluster-specific lesion maps arise naturally from clustering analysis by averaging lesion masks from patients in each cluster, rather than modeling correlations between latent variables derived from dimensionality reduction of behavioral and image variables. Third, a statistical evaluation and refinement of lesion mapping is conducted to identify lesions that distinguish behaviorally different groups. This approach assists in understanding aetiology and allows to align the complete profile of a single patient to a more general phenotype or syndrome, rather than associating a single symptom across the whole population to other single symptoms.

As mentioned in Yang et. al (2023) [21], clustering is one of the most useful methods for analyzing patient similarities for precision medicine. It groups patients into clinically meaningful subsets, which can be used for a variety of tasks, such as personalized therapies and policy making. Kim et. al (2022) [22] stress that finding similar clusters among stroke patients can be helpful from a medical perspective, as it may lead to the discovery of new patterns and more effective ways to manage stroke.

Overall, clustering is an effective method to associate many patients with each other, and to a few behavioral syndromes; to associate behavioral syndromes to specific lesion locations.

2 Materials and Methods

2.1 Data

The NIHSS data consist of 15 items describing the health state, abilities, and cognitive functions of patients. Some items are based on exterior observations that assess damage in specific areas of the brain, while other items test patients with more complex tasks, possibly involving disparate brain regions. NIHSS scores here refer to the acute phase. Statistics of NIHSS items are computed on first-stroke ischemic patients from Padua University Hospital and from Washington University in St. Louis data sets (total $n = 308$), while patients clustering is performed on the subset with both NIHSS> 0 ($n = 248$) and available brain scans ($n = 172$). While the latter enables symptoms-to-lesions mapping that would otherwise be impossible, the initial selection implies that a total NIHSS score of 0 results from lesions that do not affect function in a measurable way. The average time between the stroke event and the NIHSS assessment for the $n = 189$ ischemic subjects in Padua is $\mathrm{M} = 3, \mathrm{SD} = 3$ days, while for the $n = 119$ ischemic subjects in St. Louis is $\mathrm{M} = 13, \mathrm{SD} = 5$ days. The Level Of Consciousness (LOC)-Vigilance item is removed, since the only subject not scoring 0 is filtered out, to exclude incomplete results from unresponsive patients.

Imaging data come from CT and MRI scans, co-registered in MNI152 standard space [23] at $1\mathrm{mm}^3$ resolution, with lesions segmented by professionals. Combining cohorts and filtering, the resulting data set comprises 172 patients, whose ages ranged from 21 to 94 years (M=64, SD=15 years). A summarizing scheme of data selection is shown in Figure 1.

2.2 Correlations and Distances

Denoting P the set of n patients, and $X_{n=308,m=14}$ the matrix of NIHSS observations considered, the relationship between NIHSS items across the population is explored

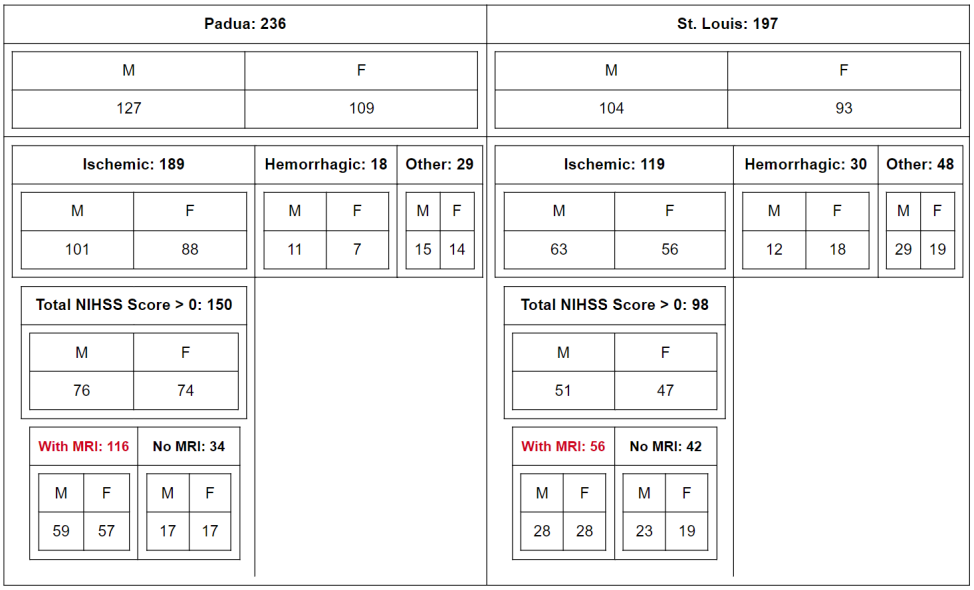

Fig. 1 Patient selection workflow. Red marks patients taken into account for clustering. Abbreviations: M: male, F: female.

with statistics suitable for ordinal data. Co-occurrences of deficits are counted irrespective of severity: for each couple of items, the number of patients exhibiting both deficits is counted. Then Spearman's rank correlation coefficient among items is computed, to measure positive and negative associations. These are all operations of type $X^T X_{(m=14,m=14)}$.

Subsetting P to the n patients with NIHSS$>$ 0 and available brain scans, $X_{n=172,m=14}$ is the matrix of NIHSS observations for subsequent analyses. The General Distance Measure [24] [25] (GDM) is the proximity measure quantifying how different two patients are, based on their profile composed of 14 NIHSS scores. The GDM is defined taking inspiration from Kendall's General Correlation Coefficient, in order to process continuous as well as ordinal values. It is here corrected and simplified in

notation as:

$$d_{ab} = \sum_{j=1}^{m} d_{abj} \quad \text{with} \quad d_{abj} = \frac{1}{2m} - \frac{-\sigma_{abj}^2 + \sum_{c=1, c\neq a,b}^{n} \sigma_{acj}\sigma_{bcj}}{2\left[\sum_{j=1}^{m}\sum_{c=1}^{n}\sigma_{acj}^2 \sum_{j=1}^{m}\sum_{c=1}^{n}\sigma_{bcj}^2\right]^{\frac{1}{2}}}, \tag{1}$$

and

$$\sigma_{abj} = \begin{cases} 1, \text{ if } x_{aj} > x_{bj} \\ -1, \text{ if } x_{aj} < x_{bj} \\ 0, \text{ if } x_{aj} = x_{bj} \end{cases} \tag{2}$$

with

- $a, b, c = 1, 2, ..., n$: indexes of the n subjects (stroke patients)
- $j = 1, 2, ..., m$: index of the m variables (NIHSS items)
- x_{aj}: value of the variable j for subject a (NIHSS score)
- σ_{abj}: sign of the difference between patient a and patient b for item j
- d_{abj}: distance between subjects a and b, on item j.
- d_{ab}: distance between subjects a and b.

Note that the GDM and its complement, the General Similarity Measure (GSM, of value $s_{ab} := 1 - d_{ab}$), depend on the whole cohort observations X for each pairwise computation. These are operations of type $XX^T_{(n=172,n=172)}$, thus computing similarities between patients can be seen as the dual problem with respect to the problem of computing correlations between variables. The matrix of all pairwise relationships is symmetric, describing an undirected weighted graph: it is hence called the adjacency matrix $W_{(n=172,n=172)}$ for the graph $G(P, W)$, with n nodes referring to the patients and $n(n-1)/2$ weighted edges referring to the GSM of all pairs of patients. For simplicity, W will hereon denote the matrix or the network, depending on context.

2.3 Repeated Spectral Clustering

Repeated Spectral Clustering (RSC) is a novel consensus clustering approach, defined in this paper based on spectral clustering, and aimed at dealing with the variability of results that stems from random initializations. Spectral clustering is the name for techniques based on the spectrum of the adjacency matrix's (or graph's) Laplacian [26]. Further arguments for using spectral clustering instead of other common techniques, such as hierarchical clustering and k-means, can be found in supplementary material Appendix B. Spectral clustering consists in two steps: spectral embedding, and clustering in the embedding space. The spectral embedding step yields Euclidean coordinates for the network nodes, in the space defined by the eigenvectors of the graph Laplacian. The actual clustering step can be any algorithm working in Euclidean spaces, typically k-means in embedding space [27], as in the present case. RSC is also a consensus clustering technique, with two distinct phases of spectral clustering. In the first phase there is evidence accumulation: the results of N different spectral clustering runs on W are recorded in the $(n \times n)$ consensus matrix C. In the second phase, C itself undergoes spectral clustering [28]. RSC is defined by the use of L_{sym} spectral clustering [27] for both phases, whereas other consensus clustering methods change algorithm between the two phases, or change algorithm and/or data subset in the N runs of the first phase. While the entries of W measure the similarities between any two subjects, the corresponding entries of C count how many times those any two subjects are clustered together over the N runs of the evidence accumulation phase [29]. The more times two subjects are found in the same cluster, the more evidence those subjects are actually similar and co-members of the "real" cluster. Matrix C too induces a weighted undirected graph and its spectral properties and visual aspect highlight how the subject nodes are better separable than they are in W. Defined in this way, RSC unifies the ability to find clusters in spectral embedding with the robustness of results regarding centroids initialization, thanks to the evidence accumulation stage.

As with other algorithms akin to k-means, the number of clusters k is set a priori or heuristically: here, the methodology relies on the spectral properties of C as a function of k. Overall, RSC can be interpreted as a way to extract the clustering signal C in the noisy W, averaging multiple measurements. The modules of RSC are described in Algorithm 1, Algorithm 2, and Algorithm 3 for spectral embedding, clustering and the whole RSC respectively. For further details refer to Appendix C.

Algorithm 1 spectral embedding

inputs: A, k
compute graph Laplacian from adjacency matrix A
compute k eigenvectors for the k smallest eigenvalues of Laplacian
employ eigenvectors as columns of matrix U_A
L2-normalize U_A rows, giving matrix T_A
return T_A ▷ T_A is the matrix of embeddings of A

Algorithm 2 clustering

inputs: T, k
compute clusters with k-means for the co-ordinates in T
return labels

Algorithm 3 Repeated Spectral Clustering

inputs: W, k
instantiate $C \leftarrow 0_{n,n}$
perform Algorithm 1 with (W,k)
set repetitions $N \leftarrow 1000$
while $N \neq 0$ **do**
    perform Algorithm 2 with (T_W,k)
    update C
    $N \leftarrow N - 1$
end while
perform Algorithm 1 with (C,k)
perform Algorithm 2 with (T_C,k)

2.4 Voxel-wise analysis

The application of RSC to the affinity matrix obtained from the GDM pairwise similarities, leads to the isolation of clusters. In order to evaluate how a cluster is different from the others, lesion density maps can be visually compared in Figure 3, Panel **B**. Each map is obtained averaging the binary lesion masks in the respective cluster: thus, each voxel intensity is the [0-1] normalized frequency with which that voxel is lesioned in the cluster considered; equivalently, it is the normalized frequency of patients in that cluster that have a lesion in that location. To quantitatively test how significant are differences between clusters, we further perform a voxel-wise analysis. Since the objective of the clustering algorithm is to separate clusters in the space related to NIHSS items, statistical test on the separation in these same variables would lead to inflated p-values. For this reason, the statistical analysis focuses on voxel that are lesioned with significantly different frequencies across clusters.

The analysis conducted is composed by the following steps:

1. The comparisons are done in a one vs all-but-one way; the null-hypothesis H_0 affirms that the lesion distribution in one cluster is not different from the lesion distribution of all the remaining patients.
2. From the $182 \times 218 \times 182$ available voxels in MNI-152 template, we select those that are lesioned at least 8 times, which corresponds to the 5% of the patients. This is both for a statistical reason, since these voxels carry a greater effect size, as well as a computational reason, since this selection considers only 241163 voxels.
3. The test performed is the difference between proportions with pooled variance [30]:

$$t_v = \frac{|\frac{v_i^+}{n_i} - \frac{v^+ - v_i^+}{n - n_i}|}{\sqrt{\frac{v^+}{n}(1 - \frac{v^+}{n})(\frac{1}{n_i} + \frac{1}{n - n_i})}} \tag{3}$$

where:

n_i: number of patients in cluster i.

v_i^+: number of times voxel v, is lesioned (+) considering patients in cluster i.

n: number of patients (172).

v^+: number of times voxel v, is lesioned (+) considering all the patients.

4. A permutational approach is used, where patients' cluster assignment is randomly shuffled; for each selected voxel the statistic is computed 5000 times. In this way the distribution of the statistic t_v can be built under H_0, where there is random assignment of patients lesions to clusters.
5. Due to the large number of hypotheses under test (1 for each voxel), a multiple-comparison correction is needed. Both Bonferroni and Holm's [31] methods are largely conservative, so a multistep-maxT (Westfall & Young [32]) procedure is used instead. This allows a Family Wise Error Rate (FWER) control at $\alpha = 0.01$.

Since RSC is performed with k=5, the presented methodology is applied five times. These other five comparisons are not taken into account by the max-T algorithm; consequently, a Bonferroni correction is applied. Therefore, the results are said to be overall corrected, with FWER at $\alpha = 0.05$.

2.5 Open-source software tools

Filtering data, statistical analyses, clustering, lesion frequency map computation and visualization are all performed with the Python 3 programming language, and its libraries numpy, pandas, scipy, scikit-learn, nibabel, and nilearn. Voxel-wise analysis are conducted with the R programming language, and its libraries foreach, doFuture, utils, scales, reticulate. The multi-step maxT correction is done using the code available at https://github.com/livioivil/r41sqrt10. In order to allow the usage of the presented analytic workflow by the scientific community on different datasets, code will be available at https://github.com/MedMaxLab/nihss_clustering; data will be available upon reasonable request to the authors.

3 Results

3.1 Correlations

The first set of results illustrates the co-occurrence of deficits, i.e. the contemporary presence regardless of severity scored, and the rank correlations of deficits across subjects, accounting for their severity (scores ranging from 0-2, to 0-4 for items in the NIHSS). Figure 2, Panel **A** shows the co-occurrence deficit matrix.

Facial Palsy and Dysarthria frequently co-occur in 68 patients. They also frequently co-occur with language deficits, with 45 patients exhibiting Facial Palsy and 34 showing Dysarthria. Both are linked with motor and sensory impairments on both sides of the body. Language deficits, specifically "Best Language", are often associated with right-sided motor deficits, sensory issues, and visual impairments. Inattention typically accompanies left-sided motor deficits, sensory problems, and visual impairments. Level Of Consciousness-Commands (LOC-C, ability to understand and execute a simple motor command) and "Best Gaze" deficits have fewer associated with other deficits.

Figure 2, Panel **B** shows only significant correlations after Benjamini-Hochberg FDR correction for multiple comparisons ($\alpha = 0.05$) on permutation-based p-values. Of 91 correlations, only 11 are significant after correction. The network plot reveals a strong correlation between arm and leg Motor deficits on both sides. "Best Language" shows significant correlation with LOC-Q, negative correlation with left Motor deficits (being often co-occurring with right Motor deficits). Left Motor deficits correlate significantly with Inattention, which in turn correlates with Visual deficits. Interestingly, Sensory deficits correlate significantly with Limb Ataxia.

In summary, the co-occurrence patterns and correlation network show that arm and leg Motor deficits positively covary on the same side of the body, and negatively covary on opposite sides. Language deficits may occur alone or with right Motor deficits,

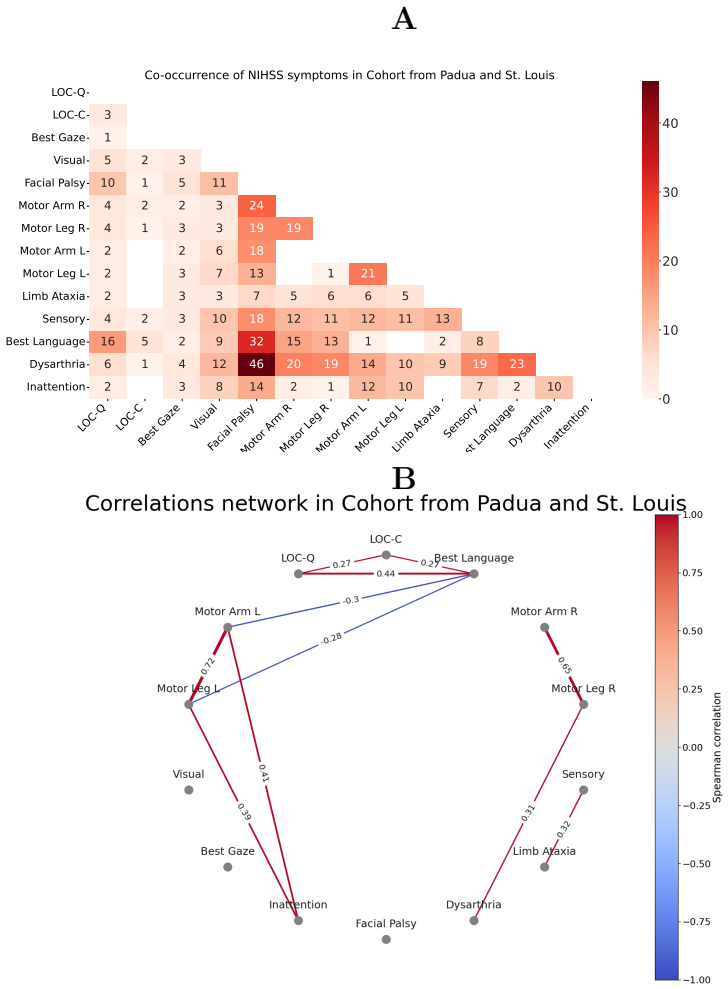

Fig. 2 (Panel **A**) Table visualization of co-occurrences of item deficits. Intense red for higher number of co-occurrences. Empty blocks represent zero co-occurrences. (Panel **B**) Graph visualization of correlation of item deficits severity. Only significant correlations ($\alpha < 0.05$, Benjamini-Hochberg FDR correction) are visualized. Abbreviations: LOC: Level Of Consciousness; LOC-C: LOC-Commands; LOC-Q: LOC-Questions; L: left; R: right; BL: Best Language; Limb-A: Limb Ataxia, Inatt: Inattention; Facial-P: Facial Palsy. Color line thickness represent the value of Spearman's ρ.

while Inattention covaries with left Motor deficits. Facial Palsy and Dysarthria are positively correlated and often occur with other deficits, regardless of side.

3.2 Repeated Spectral Clustering

Applying RSC allows to identify 5 main clusters of patients using the NIHSS data. The $k = 5$ process shows the sharpest change in the spectrum of the matrix C, as discussed in Appendix C. The minimum, median and maximum NIHSS profiles for the 5 clusters and the corresponding group average MRI lesion anatomy are visualized in Figure 3, Panel **B**. The Fruchterman-Reingold force-directed algorithm [33] is used for visualization so that patients who are more similar appear closer in the graph (see Panel **A**).

Cluster 0 includes left arm-leg Motor deficits, Inattention, Facial Palsy, and Dysarthria. In the most severe cases, Visual, "Best Gaze", and Sensory deficits can also occur (see difference between median and max deficit). The lesions localize to the vascular territory of the right middle cerebral artery (MCA).

Cluster 2 includes right arm-leg Motor deficits, Facial Palsy and Dysarthria. The lesions are roughly localized to the deep left MCA branches; in details, it involves the medial and lateral lenticulostriatal arteries of the M1 branch of the left MCA.

Cluster 4 involves selectively language deficits and Facial Palsy, and in the most severe cases can affect vision, sensation, and gaze. The lesions are localized in the territories of the superficial left MCA involving the cortical branches of the MCA, including the anterior temporal branch and the superior and inferior branches of M2.

Cluster 1 shows a high degree of deficit variability with a preference for Inattention, Visual deficits, and gaze disorders. It primarily involves the territories of the posterior cerebral artery bilaterally (both deep and superficial branches), and only partially the right MCA (bilateral posterior cerebral artery, PCA).

Finally, Cluster 3 shows a relevant presence of Facial Palsy, bilateral upper limb Motor deficits, Dysarthria and Sensory deficits. It involves the anterior choroidal arteries and penetrating branches of the MCA, Anterior Celebral Artery (ACA), and PCA bilaterally, as well as penetrating branches of the basilar artery (lacunar infarcts).

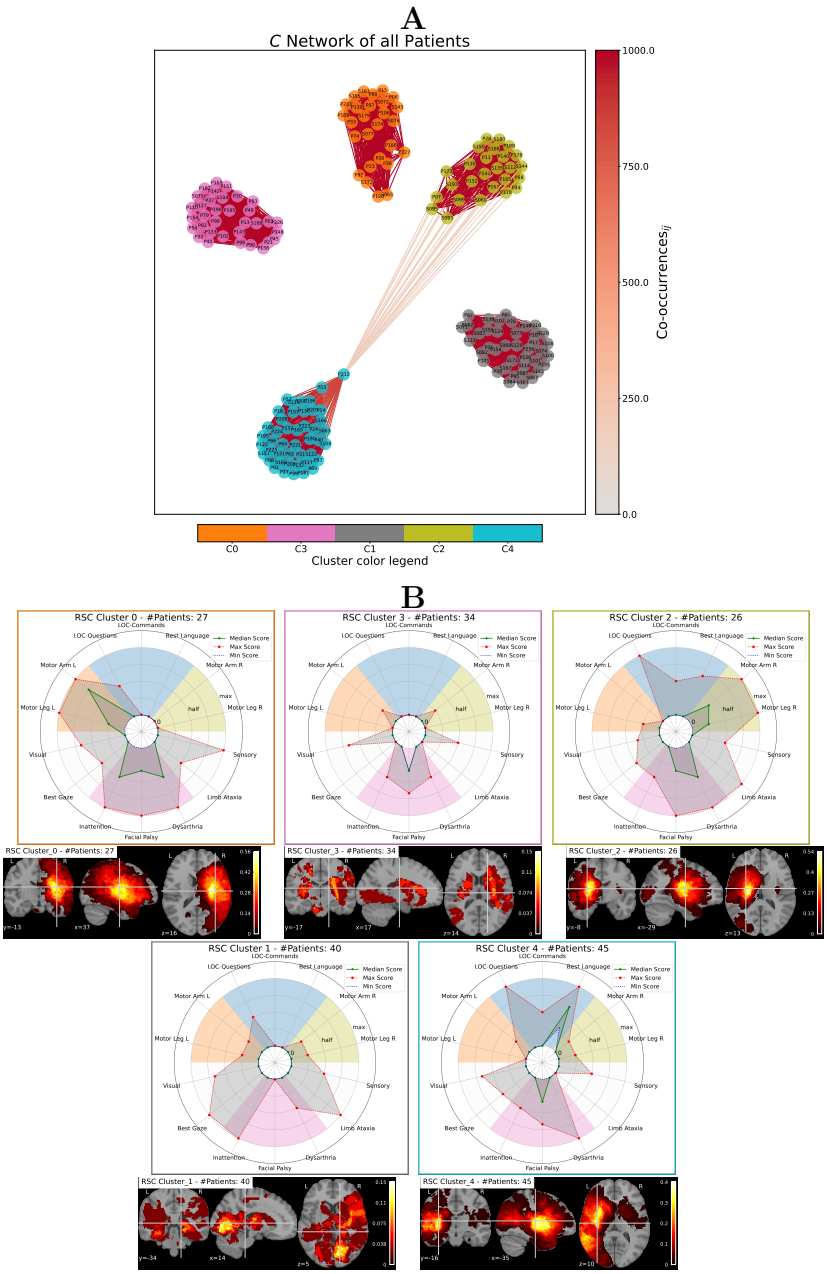

Fig. 3 (Panel **A**) Graph visualization of the co-occurrence matrix C. Nodes represent patients, while colors identifies clusters. Darker red color links patients frequently clustered together. (Panel **B**) Radar charts and lesion heat maps for the clusters obtained from RSC. In the radar plots the lines represent (in blue) the minimum, (in green) the median and (in red) the maximum of the NIHSS values inside the cluster. In the heatmaps, brighter colors indicates voxels lesioned by more subjects.

In summary, RSC provides a distribution of symptoms' clusters confirming the presence of contralateral Motor syndromes, the relative segregation of language deficits from Motor deficits, and the association of Inattention and Sensory deficits with right hemisphere lesions.

3.3 Voxel-wise analysis

Applying the voxel-wise analysis allows to find statistically significant distribution of lesions coming from three clusters, in particular Cluster 0 (left Motor deficits), Cluster 2 (right Motor deficits) and Cluster 4 (language deficits). Significant regions are shown in Figure D7. The multiple-slices views are shown in the supplementary material Appendix D.

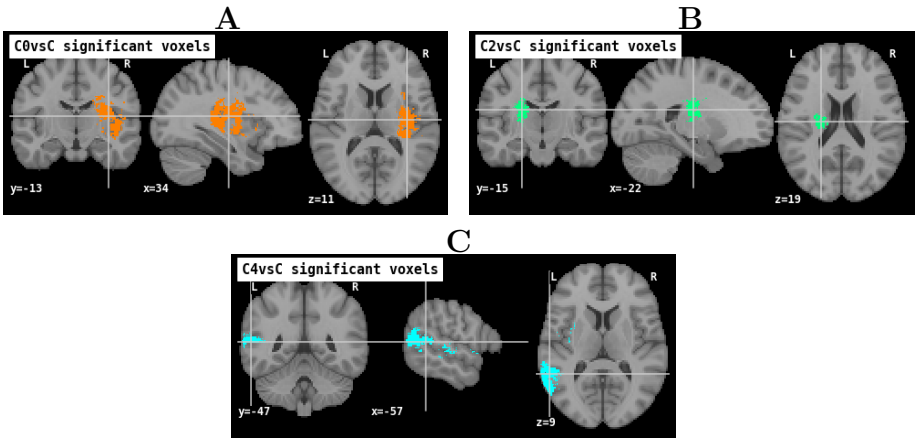

Fig. 4 Significant voxels found from the voxel-wise analysis with FWER at $p = 0.05$. At the top (Panel **A**) in orange, Cluster 0 (left motor deficits), in the middle (Panel **B**) in green, Cluster 2 (right motor deficits) and at the bottom (Panel **C**) in light blue, Cluster 4 (language deficits).

Concerning Cluster 0, the orange region in Figure D7 Panel **A**, is part of the vascular territory of the lenticulostriate arteries of the right MCA. Anatomically, this corresponds to the right corticospinal tract at the level of the corona radiata and the posterior limb of the internal capsule, as well as the posterior insula, globus pallidus, and putamen [34].

For Cluster 2, the green region in Figure D7 Panel **B**, is part of the vascular territory of the lenticulostriate arteries of the left MCA. Anatomically, this corresponds to the left corticospinal tract at the level of the corona radiata and the posterior limb of the internal capsule.

Finally, for Cluster 4, the light blue regions in Figure D7 Panel **C**, are within the vascular territory of the inferior trunk of the left M2. Anatomically, this corresponds to the posterior insula, Heschl's gyrus, superior temporal gyrus (Wernicke's area), and arcuate fasciculus [35], [36].

In summary, the patients' lesions of these three clusters are highly specific with respect to the rest of patients, and thus characteristic of these clusters.

4 Discussion

The clustering and imaging results are aligned with the medical literature. Interestingly, they add anatomical discriminative power to the NIHSS tests, leveraging co-occurrence statistics to distinguish similar deficits from different underlying lesions. Moreover, the application of RSC to the GDM/GSM measure, and the workflow code are themselves useful results, made available to researchers in biomedical domains.

4.1 The underlying structure of the NIHSS

Previous factor analyses of the NIHSS (Lyden et al. [8], Zandieh et al. [7]) identified two main factors, related to the left and right hemispheres, that can explain the majority of variance of behavioral impairment following stroke. While these results assess the conformity of the NIHSS to cerebral hemispheric lateralization, hence fundamentally validating the scale in its clinical setting, they offer limited acute stroke-phase insights for clinicians.

Here, using an unsupervised machine learning approach based on the NIHSS scores, we identify a finer-grained structure of post-stroke impairment. RSC, in addition to

left and right hemispheric clusters, identifies a group of co-occurring deficits (i.e. facial palsy, bilateral upper limb motor deficits, dysarthria and sensory deficits) and anatomical correlates (i.e. bilateral basal ganglia, internal capsules) that may be associated with lacunar syndromes. This clinical entity reflects a specific etiological mechanism (i.e. arteriolosclerosis [37]) that guides the diagnostic and therapeutic algorithms for these patients. Moreover, RSC identifies numerous left hemisphere clusters, possibly due to NIHSS's higher sensitivity to left lesions [38]. In addition, results confirm the unreliability of the limb ataxia subtest, which is scored in supra-, infra-tentorial and in bilateral lesions. This is in line with previous studies that have shown low inter-rater reliability of this subtest, proposing modified versions of the scale for clinical trials (see the mNIHSS [39], [40]).

Notably, RSC seems to identify a cluster (Cluster 1) with highly heterogeneous patients Figure 3. This cluster may be interpreted as patients without clear behavioral similarities (i.e. "leftovers") according to the NIHSS, reflecting the necessity to integrate this test with additional hyperacute clinical evaluations, including cognitive and mood assessments, as shown by previous work [9], [41].

Finally, our voxel-wise analysis identifies a statistically significant set of lesions for Cluster 0 (left Motor deficits), Cluster 2 (right Motor deficits) and Cluster 4 (language deficits). In particular, motor deficits are significantly associated with corticospinal tract damage, while language deficits with damage to the arcuate fasciculus and Wernicke's area. Interestingly, the described regions correspond precisely to anatomical substrates with a well-established prognostic role for recovery [42], [43], [44], [45], [46].

Future work, using the same validated methodological approach, could assess the longitudinal evolution and anatomical specificity of the identified clusters to better characterize post stroke recovery trajectories.

4.2 Unsupervised learning approaches in stroke

This work introduces a new perspective on stroke cohort data analysis. Patients are represented in a network where they are linked by similarity and clustered. This original view is partly shared by trajectory profile clustering [47] and latent class analysis[16] used in recovery studies. Statistical analyses have instead traditionally focused on variables' correlation, variables' covariance with imaging data, and population outcome predictions. Examples of imaging and behavioral studies can be traced to Voxel-based Lesion–Symptom Mapping (VLSM) [48] in its univariate and multivariate forms, and to the Partial Least Squares (PLS) family of applications [20]. VLSM assigns t-statistics, p-values, and correlation scores to brain regions linked with behavioral deficits, while PLS reduces and correlates dimensions of brain and behavioral data.

Overall, the traditional analyses focus on problems of the form X^TX when X is a matrix of n observations (subjects) on m variables; the matrix product is X^TY when Y has the same n subjects observed, and there are m_X variables in the set of features F_X (e.g. behavioral scores), and m_Y variables in the set of features F_Y (e.g. brain voxels). In contrast, the perspective here focuses on patient profiles, patient-to-patient and patient-to-group relations, and group-specific associations of deficits and lesions, addressing the matrix product XX^T. This explains the complementarity of this approach with existing literature. Data points are clustered on the set of features F_X (here, NIHSS sub-scores), and cluster statistics are evaluated on the feature set F_Y (binary masks of brain lesions in CT and MRI scans). The process leverages the multimodality of both F_X and F_Y, translating to different domains. Starting from first principles, the suitability of the GDM for ordinal data (F_X) is demonstrated and used to construct the similarity network of NIHSS score profiles. The most common clustering distances include Euclidean, cosine, Manhattan, Mahalanobis and Pearson distances, as noted in [49]. Differences in pairwise similarity distributions can be found

in Appendix A. Using the cosine distance would result in most patients being orthogonal, thus maximally different. However, this property is not desirable, as scoring 1 in two different items does not make two subjects maximally different, one argument being that they are both close to the healthy case where all NIHSS scores are zero. Euclidean distance is unsuitable and not theoretically motivated for the NIHSS data space, since it is a lattice space, where different distances on different axes are arguably not proportional (anisotropy and inconsistency of scale). Mahalanobis and Pearson distances are also not suitable for NIHSS data, since they assume the validity of standardization operations (mean-centering and variance-scaling) on the data, which are intrinsically skewed. Manhattan distance works well with interval data. However the GDM is still superior to Manhattan in these respects: because it accounts for the variability in each item, intensifying the impact of pairwise differences the less entropy is found in an item; because it does not depend on the magnitude of the differences, making no assumptions on the scales of different items, thus shrinking the distribution of similarities.

NIHSS items difference between two subjects conveys the difference in gravity of condition, but the gravity itself is not linear, as suggested instead by the use of single digits in the NIHSS scale. This motivates the use of GDM for such data as theoretically sound, with parsimonious assumptions, and pushing toward a more conservative position with respect to patient conditions, compared to the differences in type of symptoms. This similarity network is clustered using the unsupervised machine learning technique Repeated Spectral Clustering.

RSC unifies consensus and spectral clustering in a simple frame, taking advantage of the random initializations of k-means to derive robustness, as evidenced by the eigenvalue changes in W's and C's respective graph Laplacians. To the best of our knowledge, this is the first instance using the normalized symmetric Laplacian algorithm [27] both in the "ensemble" of random initializations constituting evidence

accumulation, and as the consensus function aggregating them in a final result, requiring a single k choice as in classical k-means, supported by the spectral analysis of W and C. For related algorithms, see [50] and [28].

Future work would extend the pipeline acknowledging its limitations, enabling its application to larger stroke patients cohorts, more extensive symptom records, and other multimodal biomedical data collections, regardless of prediction targets. While the GDM's flexibility for ordinal and continuous data is valuable, traditional preprocessing steps could support other proximity measures such as Hamming, Manhattan, and cosine distances, without hindering RSC applicability. RSC could be improved by considering geometric constraints in spectral embedding and incorporating soft partitions and probabilistic cluster attributions, currently implied by the consensus matrix. Extending these tools with a focus on interpretability can provide valuable support systems in research studies.

5 Conclusions

In recent years, machine learning and data science have advanced significantly, offering new prospects across many domains. The medical field's complexity, with its numerous variables and statistical properties, makes it difficult to create data-driven mathematical models of diseases from healthcare data. In this paper, for the first time we address stroke patients clustering, presenting a novel unsupervised data analysis pipeline tailored to the properties of health scales, and applied to a leading global cause of death and disability. Many studies have analyzed patient behavior and lesion locations, but they often rely on assumptions that can limit their validity, such as treating item modalities as numerical data and using linear models. This research complements existing viewpoints by following a new and precise methodological path and introducing key elements. Our methods focus on thorough treatment of ordinal variables and avoid restrictive and distorting assumptions, using GDM and model-free lesion mapping.

These maps show anatomical separation based on cluster membership, aligned with expected localization of deficits, despite being based solely on behavioral data. They describe a detailed structure of neurological syndromes, providing additional topographical and etiological insights for clinicians in the hyperacute phase of stroke, and confirm known limitations of widely used clinical scales.

The novel and open source workflow here provided is flexible and of high methodological value, ready for extensive adaptations to multimodal biomedical data in other domains, whenever unsupervised phenotypizing of obervations is difficult but potentially enriching, and whenever clinical healthcare data comprises ordinal scales.

Statements and Declarations

- Funding: This work was supported by the "Department of excellence 2018-2022" initiative of the Italian Ministry of education (MIUR) awarded to the Department of Neuroscience-University of Padua.
- Conflict of interest: The authors have no competing interests to declare that are relevant to the content of this article.
- Ethics approval: For data of patients of the Saint Louis cohort, this research complies with all relevant ethical regulations. Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and procedures established by the Washington University in Saint Louis Institutional Review Board. All participants were compensated for their time. All aspects of this study were approved by the Washington University School of Medicine (WUSM) Internal Review Board. For data of patients of the Padua cohort, participants from this dataset provided written informed consent following the Declaration of Helsinki principles and procedures established by the Stroke Unit and Neurological Clinic of the Hospital of Padua. This is a retrospective and non-interventional study, that

received approval from the Ethics Committee of the Azienda Ospedale Università Padova.

- Availability of data and materials: The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the contributing authors upon reasonable request. Data are located in controlled access data storages at Padova Neuroscience Center and Azienda Ospedale Università Padova.
- Code availability: Code will be available upon publication at https://github.com/MedMaxLab/nihss_clustering
- Authors' contributions: Material preparation, data collection were performed by ALB, MC, SF and LP. Data analysis was performed by MA, LFT and AZ. Medical interpretation was performed by ALB, MC and SF. Conceptualization from MA, MC, LFT and AZ. The first draft of the manuscript was written by LFT and AZ and all authors commented and edited previous versions of the manuscript. All authors read and approved the final manuscript.

References

[1] Strong, K., Mathers, C., Bonita, R.: Preventing stroke: saving lives around the world. The Lancet Neurology **6**(2), 182–187 (2007) https://doi.org/10.1016/S1474-4422(07)70031-5

[2] Feigin, V.L., Stark, B.A., Johnson, C.O., Roth, G.A., Bisignano, C., *et al.*: Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019. The Lancet Neurology **20**(10), 795–820 (2021) https://doi.org/10.1016/S1474-4422(21)00252-0

[3] Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H.: The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment. Journal of

the American Geriatrics Society **53**(4), 695–699 (2005) https://doi.org/10.1111/j.1532-5415.2005.53221.x

[4] Demeyere, N., Riddoch, M.J., Slavkova, E.D., Bickerton, W.-L., Humphreys, G.W.: The oxford cognitive screen (ocs): Validation of a stroke-specific short cognitive screening tool. Psychological Assessment **27**(3), 883–894 (2015) https://doi.org/10.1037/pas0000082

[5] Brott, T., Adams, H.P., Olinger, C.P., Marler, J.R., Barsan, W.G., Biller, J., Spilker, J., Holleran, R., Eberle, R., Hertzberg, V.: Measurements of acute cerebral infarction: a clinical examination scale. Stroke **20**(7), 864–870 (1989) https://doi.org/10.1161/01.STR.20.7.864

[6] Lyden, P.D., Lu, M., Levine, S.R., Brott, T.G., Broderick, J.: A modified national institutes of health stroke scale for use in stroke clinical trials. Stroke **32**(6), 1310–1317 (2001) https://doi.org/10.1161/01.STR.32.6.1310

[7] Ali, Z., Zahra Zeynali, K., Homa, S., Maryam, P., Sara, P., Majid, G., Mojdeh, G.: The Underlying Factor Structure of National Institutes of Health Stroke Scale: An Exploratory Factor Analysis. International Journal of Neuroscience **122**(3), 140–144 (2012) https://doi.org/10.3109/00207454.2011.633721 . PMID: 22023373

[8] Lyden, P., Claesson, L., Havstad, S., Ashwood, T., Lu, M.: Factor Analysis of the National Institutes of Health Stroke Scale in Patients With Large Strokes. Archives of Neurology **61**(11), 1677–1680 (2004) https://doi.org/10.1001/archneur.61.11.1677 . Accessed 2022-08-30

[9] Corbetta, M., Ramsey, L., Callejas, A., Baldassarre, A., Hacker, C.D., Siegel, J.S., Astafiev, S.V., Rengachary, J., Zinn, K., Lang, C.E., Connor, L.T., Fucetola, R., Strube, M., Carter, A.R., Shulman, G.L.: Common behavioral clusters and

522 subcortical anatomy in stroke. Neuron **85**(5), 927–941 (2015) https://doi.org/10.
523 1016/j.neuron.2015.02.027

524 [10] Bisogno, A.L., Favaretto, C., Zangrossi, A., Monai, E., Facchini, S., De Pel-
525 legrin, S., Pini, L., Castellaro, M., Basile, A.M., Baracchini, C., Corbetta,
526 M.: A low-dimensional structure of neurological impairment in stroke. Brain
527 Communications **3**(2), 119 (2021) https://doi.org/10.1093/braincomms/fcab119

528 [11] Cheng, B., Chen, J., Königsberg, A., Mayer, C., Rimmele, L., Patil, K.R., Gerloff,
529 C., Thomalla, G., Eickhoff, S.B.: Mapping the deficit dimension structure of the
530 national institutes of health stroke scale. eBioMedicine **87** (2023) https://doi.
531 org/10.1016/j.ebiom.2022.104425

532 [12] Woo, D., Broderick, J.P., Kothari, R.U., Lu, M., Brott, T., Lyden, P.D., Marler,
533 J.R., Grotta, J.C.: Does the national institutes of health stroke scale favor left
534 hemisphere strokes? Stroke **30**(11), 2355–2359 (1999) https://doi.org/10.1161/
535 01.STR.30.11.2355

536 [13] Fink, J.N., Selim, M.H., Kumar, S., Silver, B., Linfante, I., Caplan, L.R., Schlaug,
537 G.: Is the association of national institutes of health stroke scale scores and acute
538 magnetic resonance imaging stroke volume equal for patients with right- and
539 left-hemisphere ischemic stroke? Stroke **33**(4), 954–958 (2002) https://doi.org/
540 10.1161/01.STR.0000013069.24300.1D

541 [14] Rajashekar, D., Wilms, M., MacDonald, M.E., Schimert, S., Hill, M.D., Dem-
542 chuk, A., Goyal, M., Dukelow, S.P., Forkert, N.D.: Lesion-symptom map-
543 ping with nihss sub-scores in ischemic stroke patients. Stroke and Vascu-
544 lar Neurology **7**(2), 124–131 (2022) https://doi.org/10.1136/svn-2021-001091
545 https://svn.bmj.com/content/7/2/124.full.pdf

[15] Chen, X., Zhan, X., Chen, M., Lei, H., Wang, Y., Wei, D., Jiang, X.: The prognostic value of combined nt-pro-bnp levels and nihss scores in patients with acute ischemic stroke. Internal Medicine **51**(20), 2887–2892 (2012) https://doi.org/10.2169/internalmedicine.51.8027

[16] Sucharew, H., Khoury, J., Moomaw, C.J., Alwell, K., Kissela, B.M., Belagaje, S., Adeoye, O., Khatri, P., Woo, D., Flaherty, M.L., Ferioli, S., Heitsch, L., Broderick, J.P., Kleindorfer, D.: Profiles of the national institutes of health stroke scale items as a predictor of patient outcome. Stroke **44**(8), 2182–2187 (2013) https://doi.org/10.1161/STROKEAHA.113.001255

[17] Fischer, U., Arnold, M., Nedeltchev, K., Brekenfeld, C., Ballinari, P., Remonda, L., Schroth, G., Mattle, H.P.: Nihss score and arteriographic findings in acute ischemic stroke. Stroke **36**(10), 2121–2125 (2005) https://doi.org/10.1161/01.STR.0000182099.04994.fc

[18] Sarraj, A., Albright, K., Barreto, A.D., Boehme, A.K., Sitton, C.W., Choi, J., Lutzker, S.L., Sun, C.-H.J., Bibars, W., Nguyen, C.B., Mir, O., Vahidy, F., Wu, T.-C., Lopez, G.A., Gonzales, N.R., Edgell, R., Martin-Schild, S., Hallevi, H., Chen, P.R., Dannenbaum, M., Saver, J.L., Liebeskind, D.S., Nogueira, R.G., Gupta, R., Grotta, J.C., Savitz, S.I.: Optimizing prediction scores for poor outcome after intra-arterial therapy in anterior circulation acute ischemic stroke. Stroke **44**(12), 3324–3330 (2013) https://doi.org/10.1161/STROKEAHA.113.001050

[19] Vogt, G., Laage, R., Shuaib, A., Schneider, A.: Initial lesion volume is an independent predictor of clinical stroke outcome at day 90. Stroke **43**(5), 1266–1272 (2012) https://doi.org/10.1161/STROKEAHA.111.646570

[20] Mihalik, A., Chapman, J., Adams, R.A., Winter, N.R., Ferreira, F.S., Shawe-Taylor, J., Mourão-Miranda, J.: Canonical correlation analysis and partial least squares for identifying brain–behavior associations: A tutorial and a comparative study. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging **7**(11), 1055–1067 (2022) https://doi.org/10.1016/j.bpsc.2022.07.012

[21] Yang, W.-C., Lai, J.-P., Liu, Y.-H., Lin, Y.-L., Hou, H.-P., Pai, P.-F.: Using medical data and clustering techniques for a smart healthcare system. Electronics **13**(1) (2024) https://doi.org/10.3390/electronics13010140

[22] Kim, J.-T., Kim, N.R., Choi, S.H., Oh, S., Park, M.-S., Lee, S.-H., Kim, B.C., Choi, J., Kim, M.S.: Neural network-based clustering model of ischemic stroke patients with a maximally distinct distribution of 1-year vascular outcomes. Scientific Reports **12**(1), 9420 (2022) https://doi.org/10.1038/s41598-022-13636-w

[23] Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L.: Unbiased average age-appropriate atlases for pediatric studies. NeuroImage **54**(1), 313–327 (2011) https://doi.org/10.1016/j.neuroimage.2010.07.033

[24] Walesiak, M.: Statystyczna Analiza Wielowymiarowa W Badaniach Marketingowych. Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław, Poland (1993)

[25] Jajuga, K., Walesiak, M., Bak, A.: On the general distance measure. In: Schwaiger, M., Opitz, O. (eds.) Exploratory Data Analysis in Empirical Research, pp. 104–109. Springer, Berlin, Heidelberg (2003)

[26] Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007) https://doi.org/10.1007/s11222-007-9033-z

[27] Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm.

Advances in neural information processing systems **14** (2001)

[28] Nascimento, M.C.V., Toledo, F.M.B., Carvalho, A.C.P.L.F.: Consensus Clustering Using Spectral Theory. In: Köppen, M., Kasabov, N., Coghill, G. (eds.) Advances in Neuro-Information Processing. Lecture Notes in Computer Science, pp. 461–468. Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02490-0_57

[29] Fred, A.L.N., Jain, A.K.: Data clustering using evidence accumulation. In: 2002 International Conference on Pattern Recognition, vol. 4, pp. 276–2804 (2002). https://doi.org/10.1109/ICPR.2002.1047450

[30] Wild, C.J., Seber, G.A.F.: Comparing two proportions from the same survey. The American Statistician **47**(3), 178–181 (1993) https://doi.org/10.2307/2684972 . Full publication date: Aug., 1993

[31] Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6**(2), 65–70 (1979). Full publication date: 1979

[32] Westfall, P.H., Young, S.S.: Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. John Wiley and Sons, New York, NY (1993)

[33] Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Software: Practice and Experience **21**(11), 1129–1164 (1991) https://doi.org/10.1002/spe.4380211102

[34] Dimmick, S.J., Faulder, K.C.: Normal variants of the cerebral circulation at multidetector ct angiography. RadioGraphics **29**(4), 1027–1043 (2009) https://doi.org/10.1148/rg.294085730 . PMID: 19605654

[35] Khatri, R., Qureshi, M.A., Chaudhry, M.R.A., Maud, A., Vellipuram, A.R.,

Cruz-Flores, S., Rodriguez, G.J.: The Angiographic Anatomy of the Sphenoidal Segment of the Middle Cerebral Artery and Its Relevance in Mechanical Thrombectomy. Interventional Neurology **8**(2-6), 231–241 (2019) https://doi.org/10.1159/000502545

[36] Rhoton, A.: The Supratentorial Arteries vol. 51-4, pp. 53–120. Neurosurgery, Gainesville, FL (2002)

[37] Regenhardt, R.W., Bonkhoff, A.K., Bretzner, M., Etherton, M.R., Das, A.S., Hong, S., Alotaibi, N.M., Vranic, J.E., Dmytriw, A.A., Stapleton, C.J., Patel, A.B., Leslie-Mazwi, T.M., Rost, N.S.: Association of infarct topography and outcome after endovascular thrombectomy in patients with acute ischemic stroke. Neurology **98**(11), 1094–1103 (2022) https://doi.org/10.1212/WNL.0000000000200034

[38] Vitti, E., Kim, G., Stockbridge, M.D., Hillis, A.E., Faria, A.V.: Left hemisphere bias of nih stroke scale is most severe for middle cerebral artery strokes. Frontiers in Neurology **13** (2022) https://doi.org/10.3389/fneur.2022.912782

[39] Meyer, B.C., Hemmen, T.M., Jackson, C.M., Lyden, P.D.: Modified national institutes of health stroke scale for use in stroke clinical trials. Stroke **33**(5), 1261–1266 (2002) https://doi.org/10.1161/01.STR.0000015625.87603.A7

[40] Meyer, B.C., Lyden, P.D.: The modified national institutes of health stroke scale: its time has come. International Journal of Stroke **4**(4), 267–273 (2009) https://doi.org/10.1111/j.1747-4949.2009.00294.x . PMID: 19689755

[41] Bisogno, A.L., Franco Novelletto, L., Zangrossi, A., De Pellegrin, S., Facchini, S., Basile, A.M., Baracchini, C., Corbetta, M.: The oxford cognitive screen (ocs) as an acute predictor of long-term functional outcome in a prospective sample of stroke

patients. Cortex **166**, 33–42 (2023) https://doi.org/10.1016/j.cortex.2023.04.015

[42] Price, C.J., Seghier, M.L., Leff, A.P.: Predicting language outcome and recovery after stroke: the ploras system. Nature Reviews Neurology **6**(4), 202–210 (2010) https://doi.org/10.1038/nrneurol.2010.15

[43] Hope, T.M.H., Parker Jones, O., Grogan, A., Crinion, J., Rae, J., Ruffle, L., Leff, A.P., Seghier, M.L., Price, C.J., Green, D.W.: Comparing language outcomes in monolingual and bilingual stroke patients. Brain **138**(4), 1070–1083 (2015) https://doi.org/10.1093/brain/awv020

[44] Sophie, R., Rachel M., B., Louise, L., Hayley, W., Kate, L., Storm, A., Diego L., L.-P., Andrea, G.-V., Alexander P., L., Thomas M. H., H., David W., G., Jennifer T., C., Cathy J., P.: Better long-term speech outcomes in stroke survivors who received early clinical speech and language therapy: What's driving recovery? Neuropsychological Rehabilitation **32**(9), 2319–2341 (2022) https://doi.org/10.1080/09602011.2021.1944883 . PMID: 34210238

[45] Gajardo-Vidal, A., Lorca-Puls, D.L., team, P., Warner, H., Pshdary, B., Crinion, J.T., Leff, A.P., Hope, T.M.H., Geva, S., Seghier, M.L., Green, D.W., Bowman, H., Price, C.J.: Damage to Broca's area does not contribute to long-term speech production outcome after stroke. Brain **144**(3), 817–832 (2021) https://doi.org/10.1093/brain/awaa460

[46] Koch, P.J., Rudolf, L.F., Schramm, P., Frontzkowski, L., Marburg, M., Matthis, C., Schacht, H., Fiehler, J., Thomalla, G., Hummel, F.C., Neumann, A., Münte, T.F., Royl, G., Machner, B., Schulz, R.: Preserved corticospinal tract revealed by acute perfusion imaging relates to better outcome after thrombectomy in stroke. Stroke **54**(12), 3081–3089 (2023) https://doi.org/10.1161/STROKEAHA.123.044221

[47] Krishnagopal, S., Lohse, K., Braun, R.: Stroke recovery phenotyping through network trajectory approaches and graph neural networks. Brain Informatics **9**(1), 13 (2022) https://doi.org/10.1186/s40708-022-00160-w

[48] Bates, E., Wilson, S.M., Saygin, A.P., Dick, F., Sereno, M.I., Knight, R.T., Dronkers, N.F.: Voxel-based lesion–symptom mapping. Nature Neuroscience **6**(5), 448–450 (2003) https://doi.org/10.1038/nn1050

[49] Kumar, V., Chhabra, J.K., Kumar, D.: Impact of distance measures on the performance of clustering algorithms. In: Mohapatra, D.P., Patnaik, S. (eds.) Intelligent Computing, Networking, and Informatics, pp. 183–190. Springer, New Delhi (2014)

[50] Yu, Z., Li, L., You, J., Wong, H.-S., Han, G.: Sc^3: Triple spectral clustering-based consensus clustering framework for class discovery from cancer gene expression profiles. IEEE/ACM Transactions on Computational Biology and Bioinformatics **9**(6), 1751–1765 (2012) https://doi.org/10.1109/TCBB.2012.108

[51] Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research **3**, 583–617 (2002) https://doi.org/10.1162/153244303321897735

Appendix A Distances

Five different distances have been used to calculate the distributions of the pair-wise similarities, together with GDM. These five distances are the most used in clustering as discussed in [49]. Since Manhattan and Euclidean distances are not normalized in the 0-1 range, the pair-wise distance have been normalized respect the maximum distance found in the dataset.

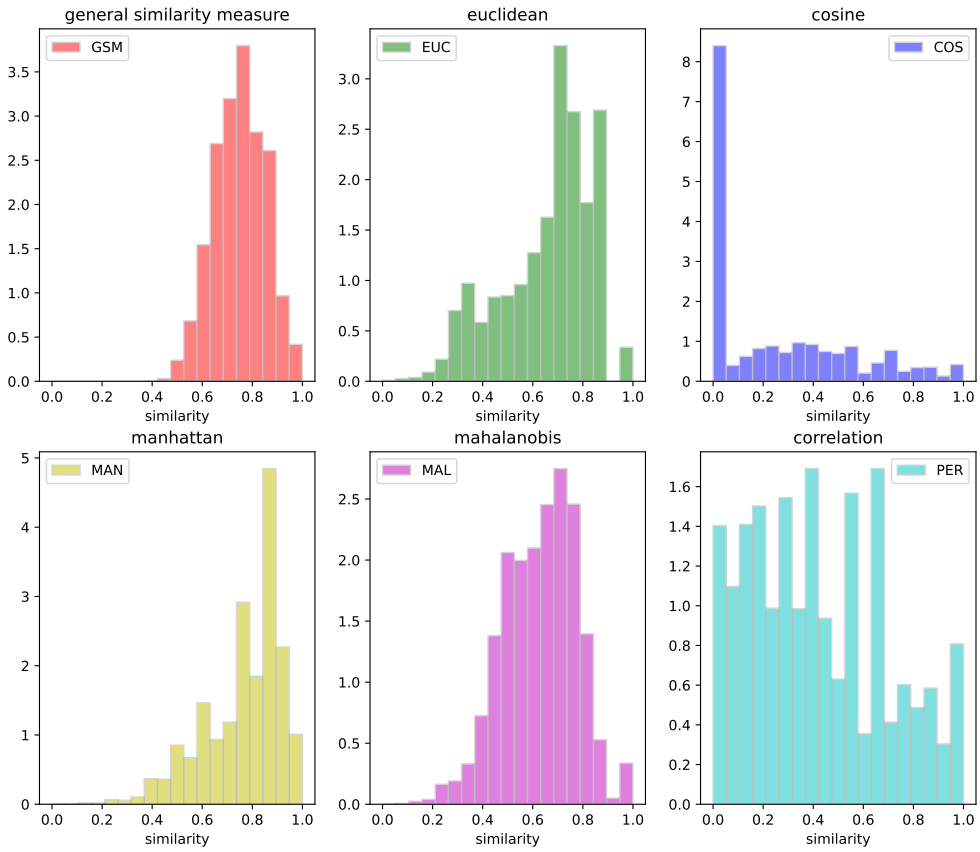

Fig. A1 Pair-wise similarities distributions, for six different distances used. GSM stays for General Similarity Measure; EUC stays for Euclidean; MAN stays for Manhattan; MAL stays for Mahalanobis; PER stays for Pearson.

Appendix B Other Clustering Techniques

k-means and hierarchical clustering are two widely used clustering techniques. This section demonstrates why these methods are unsuitable for NIHSS items and suggests alternative approaches.

Since k is a hyper-parameter of k-means, we experimented with different values ranging from 2 to 9. The raw NIHSS can be input to k-means, along with its normalized and standardized versions. The algorithm defaults to using Euclidean distance for this analysis. Concerning the *sklearn* 'KMeans' function used, the initialization was set to

random, inertia to 10, and the maximum number of iterations to 300, with a random state of 12345. Table B1 reports the number of subjects for each cluster across various normalization combinations and values of k.

k	Standardized	Normalized	Raw
2	157, 15	155, 17	156, 16
3	15, 13, 144	142, 13, 17	16, 13, 143
4	112, 15, 13, 32	111, 17, 13, 31	112, 16, 32, 12
5	93, 19, 13, 32, 15	75, 14, 39, 31, 13	37, 52, 16, 54, 13
6	13, 13, 16, 106, 19, 5	37, 13, 14, 64, 13, 31	58, 7, 9, 54, 12, 32
7	27, 13, 9, 7, 5, 19, 92	10, 18, 33, 31, 62, 5, 13	38, 21, 10, 6, 17, 67, 13
8	79, 18, 15, 5, 1, 15, 32, 26	37, 12, 10, 62, 23, 5, 10, 13	12, 12, 38, 52, 16, 32, 6, 4
9	27, 18, 15, 5, 1, 15, 13, 26, 52	37, 12, 10, 62, 23, 4, 10, 13, 1	12, 12, 38, 52, 12, 8, 5, 4, 29

Table B1 Number of subjects, for each cluster, for all the possible combinations of k and normalization used. In green the configurations with a number of subjects matching the 50%-5% criterion.

To evaluate cluster quality, consider cluster size: both very large clusters (over 50% of subjects) and very small clusters (under 5% of subjects) are problematic. Large clusters obscure phenotypic differences, while small clusters lack statistical power. Configurations with fewer than 8 or more than 85 subjects should be discarded. With these criteria, the only suitable configurations are normalized data with $k = 5$ and $k = 6$, and raw data with $k = 5$. In these configurations, clusters for Motor Right, Motor Left, and Language deficits were identified.

In all these three configurations the patients in Motor-R are 13, while the patients in Motor-L are respectively 14 for norm data and 16 for raw data; our clusters accounts better for patients with moderate and low scores in these items, giving 27 and 26 subjects respectively. The patients that are excluded from these three clusters, are 74 in our work, while 114 with k-means. The main reasons for the observed behavior could be the high-dimensionality of the data, and the spherical-shape of clusters of the k-means algorithm.

Hierarchical (agglomerative) clustering is another popular technique used to progressively aggregate data together, based on a distance measure. At each iteration, the subjects below a distance threshold are grouped together. The algorithm results in a

dendrogram, tree-like diagram in which every data point corresponds to the leafs, and branches track the merging of subjects in larger groups. Varying the distance threshold, the tree gives different numbers of clusters from a maximum equal to the number of patients, to only one clusters of all patients. The standard technique to assess the "true" number of clusters is to check the longest interval in terms of distance threshold (GDM), for which the number of clusters does not change. The longer the interval, the stronger the evidence for a structure and number of clusters. In our case, the distance is given by the GDM matrix (thus pre-computed) and the linkage mode was set to complete. Figure B2 shows the number of clusters respect the GDM distance, where the optimal k was found to be 4. However following the rules before, there are the usual right/left motors and language clusters, but cluster 0 has 96 patients, that is more than half of the dataset. In conclusion, vanilla k-means and hierarchical clustering, are not suited for the NIHSS data of our two cohorts; spectral clustering is the next most common used clustering technique and it was used as base of this work.

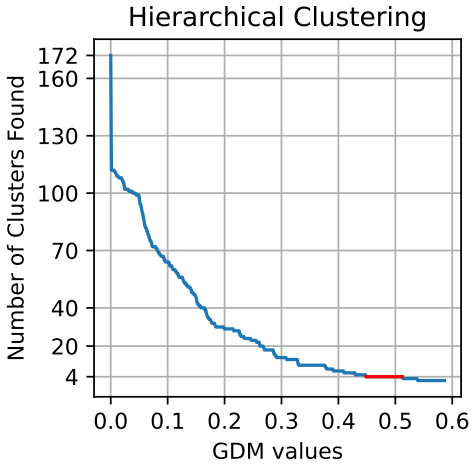

Fig. B2 Number of clusters found by the hierarchical clustering versus the distance values. In red is shown the longest interval corresponding to 4 clusters.

Appendix C Spectral Clustering

Spectral clustering is a very powerful clustering technique, however its geometrical foundation makes it hard to understand at first glance for non-experts in the field. In Figure C3 it is reported a brief illustrated example of how spectral clustering works.

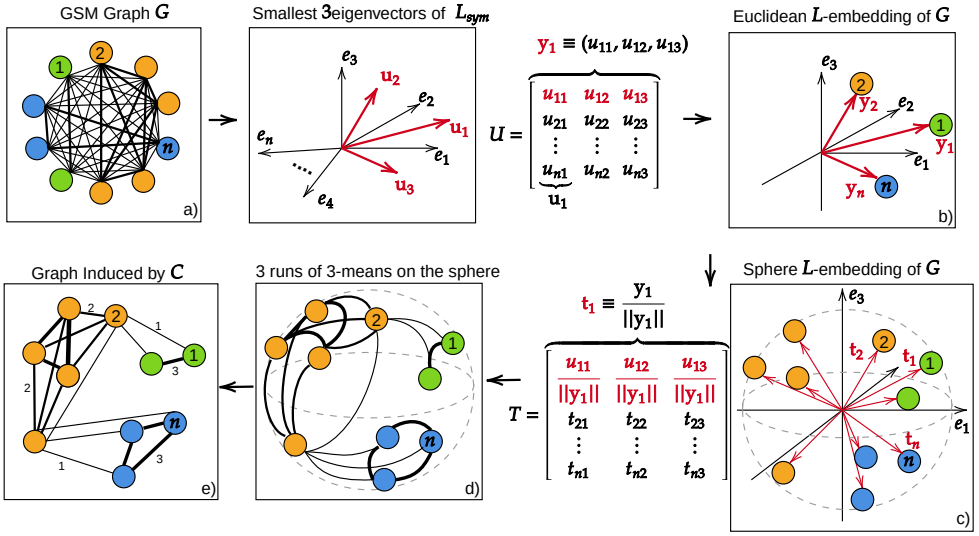

Fig. C3 Graphical summary of how spectral clustering works in the case of $k = 3$. Starting from **a)**, the GSM matrix visualized as a graph, **b)** then the matrix U is computed with its respective Euclidean embedding. **c)**Then the points are projected to the unit sphere and **d)** the k-means is run. **e)** The k-means is repeated N times giving the final co-occurrence matrix C and the corresponding graph.

From the adjacency matrix W, built from the similarity measure with $w_{ab} = s_{ab}$, the associated normalized Laplacian matrix L_{sym} is calculated [27]. If the graph of W has n separate sub-graphs, the Laplacian has n eigenvalues equal to 0. Similarly, if the Laplacian has n relatively small eigenvalues, the graph of W has n components far more connected internally than with one another. The eigenvectors associated to the n first, smallest eigenvalues are then used to create the new embedding space, where the data set is projected to. Once the spectral embedding is performed, k-means clustering is typically run over the new data set with k equal to the number n of eigenvectors

considered. k-means clustering results in a complete, hard partition of the data. The above steps would complete one level of spectral clustering. However, it is well known that k-means results depend heavily on the random initialization of the centroids. Consequently, different runs of the algorithm will yield different results and partitions. Ensembling of clustering results, also known as consensus clustering [51], is the process of aggregating different partitions from multiple algorithms into a single, robust one. Evidence Accumulation Clustering [29] is a particular consensus clustering technique. In particular, this technique is based on repeating N times the k-means algorithm. Through repetitions, evidence accumulates regarding which data samples consistently and reliably cluster together. The so called co-occurrence matrix C is defined, where the entries C_{ab} are the number of times two subjects a and b are clustered together over the N trials, as shown in Figure C4.

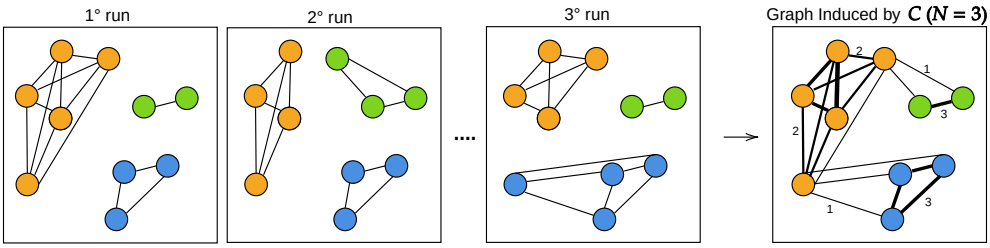

Fig. C4 Toy example of how the C matrix is built. Here different data points are clustered together differently for each run of the k-means. At the end the C matrix is obtained by counting the number of times a link is present between two nodes over N runs.

The matrix C describes a new graph, where strong connections represent frequent associations in the same clusters. The default hard partition from k-means is lost to a soft partition of connected communities. Since the objective of clustering is still assigning labels to data points, a further clustering method could be applied to the co-occurrence matrix C, to finally separate the communities from the new co-occurrence network. In this work a further spectral embedding and k-means is performed on C, given that its spectral properties and separability are enhanced compared to the

original adjacency matrix W. This idea is analogous to work in [28], where the spectral embedding relies on a different Laplacian form, L_{rw}, and the clustering for evidence accumulations are several and different algorithms. The way in which n is chosen for W and C relies on the spectrum of the normalized Laplacian of the matrix C. As stated above, n null eigenvalues indicate n completely separate groups in the data graph, and n very small eigenvalues followed by a larger eigenvalue indicate n softly distinct groups. So, in the absence of exactly 0-valued eigenvalues, the problem consists in finding a spectral gap, i.e. a large difference between an eigenvalue and all the other quasi-0 predecessors. The spectral-gap is measured as the difference between the k-th and the $k+1$-th ordered eigenvalues of the Laplacian matrix of C. At different values of n, k when working on W, one can end up with different spectral gaps in the spectrum of C. In Figure C5 is reported the spectral-gap profile for different n, k trials. As can be seen, for $k = 5$ there is the highest values of the spectral gap: by the max-gap criterion it is the suggested choice.

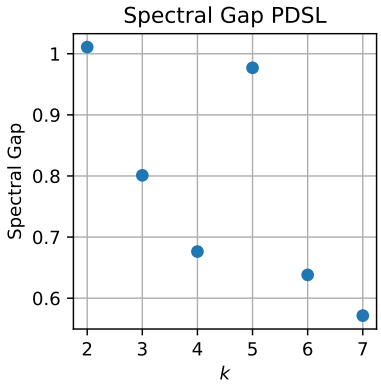

Fig. C5 Spectral-gap profile for various k, for the Padua and St. Louis dataset together.

C.1 Statistics

Considering the number of times N the k-means has to be executed, the event *two patients are clustered together* is a dichotomous event; hence C_{ab} follows a Binomial

distribution. Since C can be written as $C = NP$, C_{ab} are interpreted as mean values, so the error associated to each entry scales as $\frac{1}{\sqrt{N}}$. From statistical reasoning on the error of the variance, the minimum number of trials is therefore 200; in this work $N = 1000$. Going into the details of the k-means function implemented in *scikit-learn*, there is a parameter called n_{init} that indicated how many runs have to be executed to give the final answer, i.e. the labels. The idea behind is to try different runs and choose that with smallest inertia, as best candidate out of the batch. Further details can be seen in the corresponding web page https://scikit-learn.org/stable/. Since in this work the k-means is repeated several times in order to fully exploit the stochasticity of the centroids initialization, in principle $n_{init} = 1$. However as can be seen in Figure C6, moving from $n_{init} = 4$ to $n_{init} = 1$ gives rise to a bell-like shape probability distribution function, centered in 54; still the main peak of the inertia is in the bins $[42, 43]$. This indicates that increasing n_{init} a little bit (e.g. 4) is beneficial

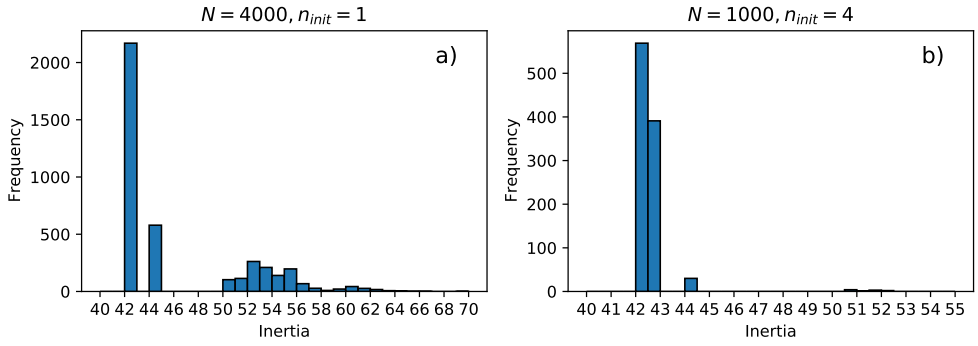

Fig. C6 Comparison between histograms of the inertia in the case $n_{init} = 1$ a) and $n_{init} = 4$ b). The total number of runs is thus fixed to $N * n_{init} = 4000$.

for having clusters with low inertia and avoid to have noisy-configurations with inertia around 54. However even in the case with $n_{init} = 1$ the final labels are equal to the case of $n_{init} = 4$.

Appendix D Voxel-wise Results

The statistical significant lesions shown in Figure D7, are here visualized at different slices.

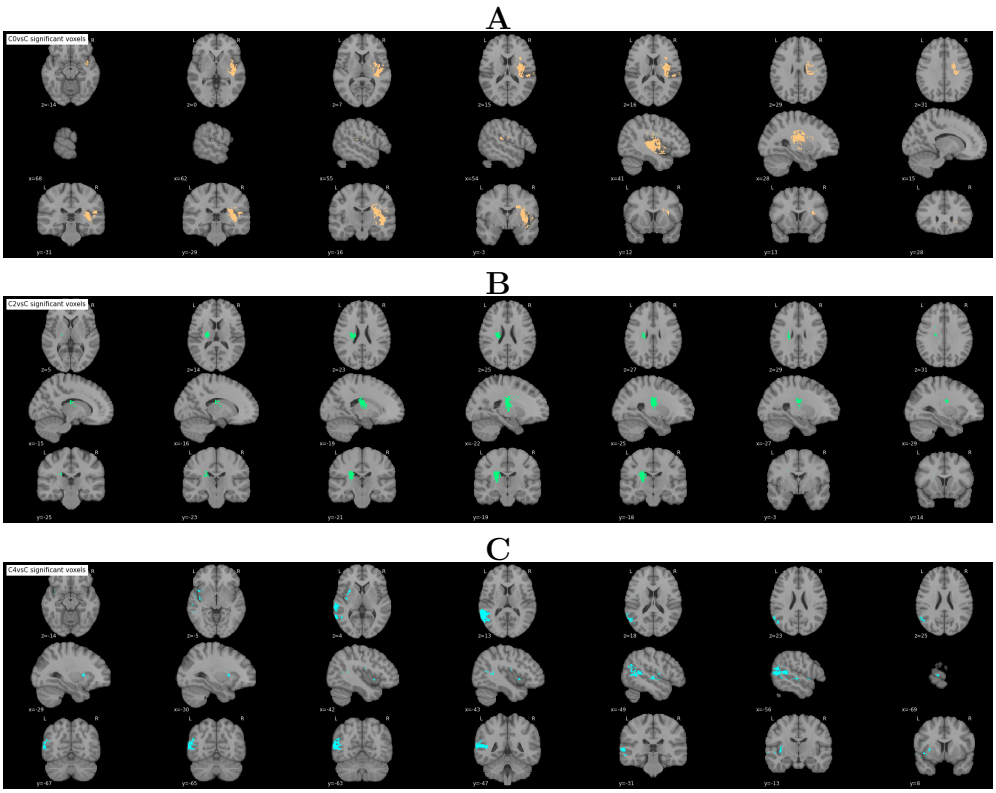

Fig. D7 Significant voxels found from the voxel-wise analysis with FWER at $p = 0.05$ for different slices. At the top (Panel **A**) in orange, Cluster 0 (left motor deficits), in the middle (Panel **B**) in green, Cluster 2 (right motor deficits) and at the bottom (Panel **C**) in light blue, Cluster 4 (language deficits).