Behavior Clusters in Ischemic Stroke using NIHSS

Louis Fabrice Tshimanga[1,2], MS
louisfabrice.tshimanga@unipd.it

Andrea Zanola[1,2], MS
andrea.zanola@phd.unipd.it

Silvia Facchini[1], PhD
silvia.facchini@unipd.it

Antonio Luigi Bisogno[1,2], MD
antonioluigi.bisogno@studenti.unipd.it

Lorenzo Pini[1,3], PhD
lorenzo.pini@unipd.it

Manfredo Atzori[1,4], PhD
manfredo.atzori@unipd.it

Maurizio Corbetta[1,2,3], MD
maurizio.corbetta@unipd.it

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

Abstract

BACKGROUND Stroke is one of the leading causes of death and disability. The resulting behavioral deficits can be measured with clinical scales of motor, sensory, and cognitive impairment. The most common of such scales is the National Institutes of Health Stroke Scale, or NIHSS. Computerized tomography (CT) and magnetic resonance imaging (MRI) scans show predominantly subcortical or subcortical-cortical lesions, with pure cortical lesions occurring less frequently. While many experimental studies have correlated specific deficits (e.g. motor or language impairment) with stroke lesion locations, the mapping between symptoms and lesions is not straightforward in clinical practice. The advancement of machine learning and data science in recent years has shown unprecedented opportunities even in the biomedical domain. Nevertheless, their application to medicine is not simple, and the development of data driven methods to learn general mathematical models of diseases from healthcare data is still an unsolved challenge.

METHODS In this paper we measure statistical similarities of stroke patients based on their NIHSS scores, and we aggregate symptoms profiles through two different unsupervised machine learning techniques: spectral clustering and affinity propagation.

RESULTS We identify clusters of patients with largely overlapping, coherent lesions, based on the similarity of behavioral profiles .

CONCLUSIONS Overall, we show that an unsupervised learning workflow, open source and transferable to other conditions, can identify coherent mathematical representations of stroke lesions based only on NIHSS data.

Keywords: Stroke, NIHSS, Spectral Clustering, Affinity Propagation, Unsupervised Machine Learning, Neuroimaging.

Non-standard Abbreviations and Acronyms

AP Affinity Propagation.

GDM General Distance Measure.

GSM General Similarity Measure.

NIHSS National Institutes of Health Stroke Scale.

RSC Repeated Spectral Clustering.

1 Introduction

Stroke is a significant global cause of death, and impairment [1] resulting from disruptions in brain blood flow. Strokes are caused by sudden decrements of flow to an area, due to artery obstruction (ischemic) or rupture (hemorrhagic). Restoring flow is critical for stroke treatment, while understanding the relationship between behavioral deficits and damage is essential for diagnosis and rehabilitation. The study of sensory-motor and cognitive impairments post-stroke has traditionally provided valuable insights into the functional organization of the brain. Historical observations of patients with specific language deficits have established the existence of specialized cortical regions such as Broca and Wernicke areas. However, it has later become clear that, on one hand, cognitive functions are not carried out by individual regions but networks of interacting regions, and, on the other, strokes do not cause individual deficits but sets of correlated deficits recurring across patients. These behavioral phenotypes are therefore important both theoretically to understand the interaction among cognitive functions, and clinically to track impairment and the efficacy of treatment. While measurements of behavioral deficits can take a lot of time using neuropsychological scales, in the clinic a rapid assessment of impairment is necessary.

The National Institutes of Health Stroke Scale (NIHSS) is the most commonly adopted clinical scale in stroke, and it is widely used to measure initial impairment and response to acute interventions. Several studies have employed a principal component analysis (PCA) of the NIHSS to describe the correlation of individual deficits across subjects. Zandieh [2] and Lyden et al. [3] described two factors: the first related to lesions in the left hemisphere and the second to lesions in the right hemisphere. These two factors, each split into motor and cognitive, explain about 80% of behavioral variance across subjects. Using a more in-depth assessment, Corbetta et al. [4] described three components, one related to language and memory, and two related to hemisphere-specific (left, right) motor and attention deficits. Bisogno et al. [5] confirmed this three component structure using the NIHSS in combination with the Oxford Cognitive Screen.

In this study, we identify robust clusters of stroke patients solely based on the NIHSS scores, through two different unsupervised learning techniques. We employ descriptive statistics to analyze the different aspects of behavior measured by the NIHSS. Profiles of subjects are compared, and their grade of difference is quantified using the General Distance Measure [6] (GDM). Similarities computed from the distances allow for a mathematical

representation of patients through a network model. We use two clustering approaches: Repeated Spectral Clustering (RSC) and Affinity Propagation [7] (AP), which natively handle network data. RSC and AP are fundamentally different. RSC extends Spectral Clustering, a graph-based algorithm requiring the number of clusters to be set in advance, with Evidence Accumulation Clustering [8], where clustering algorithms are repeated to assess robust distance measures between samples. While RSC generates a latent space from the input network, AP enables the network to reconnect dynamically into smaller units. AP is a message-passing algorithm and it does not require guessing the number of clusters, which is determined according to data. Validating similar patient groups and syndromes using different approaches confirms the intrinsic data properties captured by the two algorithms. To the best of our knowledge, these clustering techniques, along with the GDM, have not been previously employed in this field. The validity of both approaches is further confirmed by computing cluster-wise density maps of brain lesions segmented from CT and MRI scans.

2 Methods

The methods in this work rely on two complementary perspectives. First, NIHSS items are modelled as variables, whose correlation structure is analysed. Second, patients are treated as variables, whose correlation structure is converted into a network model. The network model is clustered into groups of similar patients, following two different algorithms. Resulting clusters are compared across the two algorithms, and compared against their lesion frequency density maps. 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. All of these tools are open source, code will be available at `https://github.com/MedMaxLab/nihss_clustering`; data will be available upon reasonable request to the authors.

2.1 Dataset

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 NIHSS> 0 and available brain scans (n=172). Imaging data come from CT and MRI, coregistered in MNI152 standard space [9] at 1mm^3 resolution, lesions segmented by professionals. Combining cohorts and filtering, the resulting data set comprises 172 patients, age spanning from 21 to 94 years (mean of 64 $\pm$ 15, median of 64 years). A summarizing scheme of data selection is shown in Figure 1. The NIHSS data consist of 15 items describing health state, abilities, cognitive functions of patients. Some items assess damages in specific areas of the brain, while others test patients with more complex tasks, possibly involving disparate brain regions. NIHSS scores here refer to the acute phase. The Level Of Consciousness (LOC)-Vigilance item is removed, since the only subject not scoring 0 is filtered out, so that results refer only to keenly responsive patients.

2.2 Co-occurrences, correlations between variables

Statistics are computed for the set of 308 ischemic patients. The item correlation analysis is related to PCA. Instead of analysing covariance of continuous scores with PCA, items are hereby maintained as ordinal variables.

Pairwise co-occurrences count whenever two deficits appear (score > 0) in the same patient, regardless of severity. Pairwise correlations between item scores are computed with the Spearman's rank correlation coefficient ρ, given the ordinal nature of data. Compared to PCA, Spearman's ρ and its p-value are computed independently of the distribution or scale of variables, p-values are computed via permutation test (100000 permutations) and corrected for multiple comparisons with the Benjamini-Hochberg procedure, with significance level $\alpha = 0.05$.

2.3 Correlation, similarity between subjects

We compute all pairwise distances and similarities for the patients' NIHSS profiles, filling double-entry tables. The General Distance Measure, based on Kendall's General Correlation Coefficient, is defined in [6] for this purpose. Notation is hereby simplified as:

$$d_{ab} = \frac{1}{2} - \frac{\sum_{j=1}^{m} -d_{abj}^2 + \sum_{j=1}^{m} \sum_{c=1, c\neq a,b}^{n} d_{acj} d_{bcj}}{2\left[\sum_{j=1}^{m} \sum_{c=1}^{n} d_{acj}^2 \sum_{j=1}^{m} \sum_{c=1}^{n} d_{bcj}^2\right]^{\frac{1}{2}}}, \quad \text{where} \quad d_{abj} = \begin{cases} 1, \text{ if } x_{aj} > x_{bj} \\ -1, \text{ if } x_{aj} < x_{bj} \\ 0, \text{ if } x_{aj} = x_{bj} \end{cases} \tag{1}$$

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)

d_{abj}: sign of the difference between patient a and patient b for item j

d_{ab}: distance between subjects a and b.

In summary, d_{ab} accounts for how many times patients a and b outscore one another, and the rest of the cohort (c's), on the several NIHSS items. The explicit similarity employed in this study is defined as

$$s_{ab} = 1 - d_{ab} \tag{2}$$

so to lie in the $[0, 1]$ interval (0 for "no similarity" and 1 for "complete similarity"), and it is hereon referred to as General Similarity Measure (GSM). Thus, the matrix of pairwise similarities between patients can be naturally treated as an adjacency matrix, i.e. the matricial description of a network. We call W the GSM matrix, network, graph interchangeably, depending on the aspects to highlight, where the entry in row a, column b is $w_{ab} = s_{ab}$. Network nodes represent patients, network edges represent similarity between nodes: the a, b-th entries of the matrix record the weight of the connection, which is the similarity between a-th and b-th patients. Note that these pairwise measures depend on all the sampled subjects.

2.4 Repeated Spectral Clustering

The structure of patients relations is explored using clustering techniques suited for network data. The graph model helps recover possible clusters of unconventional shapes and densities. Spectral clustering is a promising

technique that uses the set of eigenvalues (spectrum) of a graph's adjacency matrix, here the GSM matrix W. It consists in two steps: spectral embedding, and proper clustering [10]. Spectral embedding projects data points from the network representation to a latent space. Proper clustering is typically performed by the k-means algorithm in this space. However, k-means results depend on random initialization. In this work, 2-level Repeated Spectral Clustering (RSC) consists of a first level where spectral clustering is repeated N times on matrix W, and a second level where spectral clustering is applied to matrix C, the output of the first level. Repeating the k-means after the first spectral embedding allows to build the new matrix (and graph) C, where entries for patients a and b (C_{ab}) are the number of times they are clustered together, instead of their GSM. If patients a, b are often grouped together, statistical evidence accumulates [8] for them being similar. C records this evidence, and is easier to cluster than W, based on its spectrum, as in [11]. For more details on the number N please look at the Supplemental Methods. Since the graph C does not have geometric shape a priori, Fruchterman-Reingold force-directed algorithm [12] is used for visualization. In this algorithm network edges act as mechanical springs: patients often clustered together are pulled closer.

2.5 Affinity Propagation

Affinity Propagation is a clustering technique available for graphs that represent similarities between subjects [7]. The algorithm models nodes of the graph as subjects that exchange messages through channels. Patients data points communicate through the GSM network until a set of representative patients is elected, and all other subjects connect to only one of these. All other edges (similarities) are cut (set to zero). The process is iterative and stops when votes stabilize, or when the maximum number of iterations is reached. While most clustering algorithms require to specify an a priori guess for the number of clusters in the data, Affinity Propagation autonomously recovers the number of clusters ex post. Voting dynamics depends on similarity and parameters preference p, the self-vote tendency, and λ, the relative weight of the previous voting iteration on the next. In this work, the parameters are set through a grid search, balancing cluster number and in-cluster similarity.

Please see Supplemental Methods for details on mathematical formulations, parameters tuning, objective function.

2.6 Visualization of the NIHSS Clusters

Visualizing the features of NIHSS based clusters at a glance is challenging, since patients are represented by their NIHSS profiles as vectors of sub-scores. Therefore we use radar charts for clusters, paired with lesion frequency density maps (which are aimed at validating the cluster with medical knowledge). Positions of the 14 tests on the radar chart, as in the item network, are based on anatomical and correlation considerations. Symmetric motor items are on opposing sides, while the language items clique is above and Facial Palsy is below, centered. Inattention and visual items, of known and stronger association with the left clique, are closer, while Sensory and Limb Ataxia are symmetrical to them. The radius of the chart is normalized so that the maximum of the scale of each NIHSS item always lies on the circumference. Lesion frequency density maps are simple and effective visualizations of which brain areas are more often damaged. Since lesion binary masks are co-registered to the MNI152 atlas, the sum of all masks records the absolute frequency of damage for each voxel in standard space.

Summing only masks from a specific cluster and dividing by the number of subjects gives the frequency density of lesions for every voxel in said cluster. The highest values are for voxels relatively often damaged and shared in a given behavioral cluster.

3 Results

3.1 Co-occurrence and Correlation Network

We first examine two measures of deficit correlation: the co-occurrence of deficits, i.e. the contemporary presence regardless of relative severity scored, and the correlations of deficits across subjects, accounting for their severity (scores ranging 0-2, to 0-4 for items in the NIHSS). Figure 2, panel **A** shows the co-occurrence deficit matrix. Facial Palsy and Dysarthria are the most common deficits and co-occur most frequently, with $n = 68$ patients showing both deficits. Facial Palsy ($n = 45$) and Dysarthria ($n = 34$) also show the strongest co-occurrence with language deficits (NIHSS "Best Language"). Facial Palsy and Dysarthria are also associated with both right and left Motor impairments, and Sensory deficits. "Best Language" deficits occur preferably with right Motor deficits, Sensory, and Visual deficits. Inattention occurs with left Motor deficits, Sensory, and Visual deficits. Level Of Consciousness-Questions (LOC-Q, orientation to time and age) and "Best Gaze" deficits are relatively isolated. Figure 2, panel **B** shows only significant correlations after Benjamini-Hochberg procedure for multiple comparisons ($\alpha = 0.05$) on permutation-based p-values. Of 91 correlations, only 17 are significant after correction. From the network plot, a strong correlation is evident between arm and leg for both left and right Motor deficits. "Best Language" correlates significantly with LOC-Q, and right Motor deficits. Both language and right arm Motor deficits negatively correlate with left Motor deficits. Left Motor deficits show significant correlation with Inattention, which in turn correlates with Visual deficits. Interestingly, Sensory deficits show a significant correlation with left Motor deficits and Limb Ataxia. Facial Palsy correlates with Dysarthria and right Motor deficit. In summary, the patterns of deficit co-occurrence 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 of the body. Language deficits can occur in isolation or with right Motor deficits, while Inattention co-varies with left Motor and Visual deficits. Facial Palsy and Dysarthria are positively correlated, and occur frequently with other deficits irrespective of side, as they are not measured separately for the left and right sides of the body in the NIHSS protocol.

3.2 Repeated Spectral Clustering

Applying Repeated Spectral Clustering (RSC) to the NIHSS data allows to identify 5 main clusters of patients, with a small number of patients in intermediate positions among them. The Fruchterman-Reingold force-directed algorithm is used for Visualization such that patients who are more similar appear closer in the graph. The NIHSS profiles for the 5 clusters and the corresponding group average MRI lesion anatomy are visualized in Figure 3. RSC Cluster 0 ($n = 27$) 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. The mirror cluster

in the left hemisphere Cluster 2 ($n = 26$) includes right arm-left Motor deficits, Dysarthria, and Facial Palsy, and roughly localizes to the deep left middle cerebral artery (MCA) branches. Cluster 4 ($n = 45$) involves selectively language deficits and Facial Palsy, and in the most severe cases can affect vision, sensation, and gaze. The lesions are clustered in the left cortical MCA distribution (both anterior and posterior perisylvian cortex) with separate foci in anterior frontal cortex/white matter and temporal cortex. Cluster 1 ($n = 40$) shows a high degree of deficit variability with a preference for Inattention, Visual deficits, and gaze disorders. The lesions are localized posteriorly bilaterally and with a right hemisphere dominance anteriorly. Finally, Cluster 3 ($n = 34$) shows a preference for Facial Palsy, bilateral upper limb Motor deficits, Dysarthria and Sensory deficits. Lesions are subcortical and bilateral.

3.3 Affinity Propagation

In order to further explore the structures of the similarity network of patients, we also perform Affinity Propagation (AP) clustering. The algorithm results in the list of exemplars (the single nodes voted as centroids of their community) and the associations of each subject to its exemplar. After parameter tuning (see Supplemental Method) 7 clusters are obtained (Figure 4). Cluster 1 ($n = 20$) corresponds to deep right MCA lesions and includes patients with left Motor deficits. It is analogue to Cluster 0 of the Spectral Clustering (RSC). The mirror is Cluster 3 ($n = 12$) involving right Motor deficits and left MCA lesions (Cluster 2 RSC). The language deficit related Cluster 4 (RSC) splits in Cluster 0 ($n = 22$) with Facial Palsy and left perisylvian anterior lesions, and in Cluster 4 ($n = 21$) with lesions more posterior in left perisylvian temporal cortex, and Visual deficits in the most severe cases. Cluster 5 ($n = 41$) with Inattention, Visual and Sensory deficits, mainly localizes in the cortical territory of the RMCA. Two additional clusters were identified: Cluster 2 ($n = 32$) shows a high degree of variability including bilateral Motor, Visual and Sensory deficits, and Cluster 6 ($n = 24$) with Sensory, Limb Ataxia and Visual deficits related to thalamic bilateral, occipital and brainstem damage.

In summary, the two methods provide a very similar 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.

3.4 Clusters validation

In the final step, we cross-check the results of the two clustering methods. Table 1 describes all clusters in terms of item damage and vascular lesion. A comparison is shown in a Sankey plot, and described below keeping into account the difference in the number of clusters obtained (Figure 5). The reported percentage of shared patients between one RSC cluster and one AP cluster is always relative to the number of patients in the RSC cluster, since only AP clusters can happen to be completely included in RSC clusters. Cluster 0 RSC ($n = 27$) and Cluster 1 AP ($n = 20$) correspond well to deep RMCA stroke lesions, sharing 67% of patients. Cluster 1 RSC ($n = 40$) and Cluster 6 AP ($n = 24$) correspond to highly variable clusters with patients without clear behavioral similarities; however these two clusters share 47% of patients. Cluster 2 RSC ($n = 26$) and Cluster 3 AP ($n = 12$) correspond well to LMCA stroke lesions; these two clusters share 46% of patients. Cluster 3 RSC ($n = 34$) and Cluster 5 AP ($n = 41$) share 74% of patients, showing a preference for Facial Palsy. Cluster 4 RSC ($n = 56$), Cluster 0 AP

($n = 22$) and Cluster 4 AP ($n = 21$) correspond well to left perisylvian anterior lesions; the first cluster shares with the last two respectively 49% and 47% of patients, giving an overall overlap of 96% of subjects.

4 Discussion

In recent years, machine learning and data science have grown significantly, offering previously unimaginable prospects in many domains. However, their application to a challenging field like medicine is often complex and difficult. Due to factors such as the statistical characteristics and heterogeneity of patients data, it is still difficult to develop data-driven algorithms that operate generalizable mathematical models of diseases from healthcare records. As an attempt in the direction of developing procedures for unsupervised learning from medical data, this paper analyzes the clustering of stroke patients based on NIHSS scores only.

4.1 The underlying structure of the NIHSS

Previous factor analyses of the NIHSS (Lyden et al. [3]; Zandieh et al. [2]) identified two main factors, related to the left and right hemisphere, 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 do not provide additional insights for the clinician in the acute phase of stroke. Here using an unsupervised machine learning approach based on the NIHSS scores we identify a finer-grained structure of post-stroke impairment. Specifically, the Repeated Spectral Clustering technique, 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 [13]) that guides the diagnostic and therapeutic algorithms for these patients. The Affinity Propagation identifies additional clusters of patients. In particular, the co-occurrence of visual and sensory deficits, segregates left posterior from left anterior MCA lesions. Both clusters show language deficits, but only the posterior cluster is associated with visual and sensory impairments. This distinction possibly corresponds to the separation between anterior and posterior aphasia syndromes (i.e. productive/motor aphasia vs. sensory aphasia). This is particularly interesting as the NIHSS language subtest alone does not distinguish different types of aphasia. Finally, posterior circulation strokes are also identified based on common behavioral profiles (Cluster 6). The NIHSS typically underestimates this pathophysiological entity. Many authors have attempted to develop a modified version of the NIHSS (POST-NIHSS [14]) to improve prognostic accuracy. On the contrary, our results show that co-occurrence of deficits on the NIHSS alone, is able to identify brainstem, cerebellar or occipital strokes. Overall, the patients' NIHSS profile obtained from these new clustering methods was able to provide additional etiological (i.e. lacunar stroke) and topographical data (i.e. posterior circulation stroke) in the hyperacute phase of stroke. Other observations regarding the known limitations of this clinical scale are worth noting. Both unsupervised learning techniques identify a higher number of clusters in the left hemisphere. Possibly, this reflects the higher sensitivity of the NIHSS towards left lesions [15]. In addition, our 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 [16], [17]). Finally, both clustering approaches seem to identify a highly variable cluster (Cluster 1 $n = 40$; Cluster 2 $n = 32$) of patients (Figure 3 and Figure 4). This is also confirmed in the following Sankey plot analysis (Figure 5). These clusters may be interpreted as patients without clear behavioral similarities (i.e. "leftovers") according to the NIHSS, reflecting the necessity to integrate this evaluation with additional hyperacute clinical evaluations, including cognitive and mood assessments, as shown by previous work [4], [18].

4.2 Unsupervised learning approaches in stroke

Clustering holds significant potential for modeling complex syndromes. Still, challenges in applying similar methods persist due to varying outcomes in different unsupervised learning approaches. Repeated Spectral Clustering (RSC) is a powerful technique based both on geometrical and statistical considerations. RSC has the embedding capacity of the data inherited by spectral clustering and the statistical interpretation given by consensus clustering. Indeed from Figure 3 it is easy to see that some patients, which the most evident is $\{P212\}$, are not clearly clustered but are shared over repetitions, between two or more clusters. This suggests that those patients have a NIHSS profile that is similar to patients of the other clusters. Thus RSC is useful to inspect relations between clusters and subjects of difficult and/or dubious attribution. Affinity Propagation offers a complementary approach to Repeated Spectral Clustering. Although clusters are completely separated from one another, with the exclusive choice of "exemplars", each subject assignment can be explicitly inspected and evaluated comparing said subject to its exemplar, and against the others. The set of exemplars offers a context in which other patients' profiles can be understood. Moreover, changing the interpretable parameter p, or the parameter tuning process, it is possible to choose coarser or finer grained context and clustering. This study has limitations regarding the variability of outputs: namely, Affinity Propagation is sensitive to 2 hyperparameters, only one of which was designed to qualitatively affect results; on another level, results from Affinity Propagation (once settled) and Repeated Spectral Clustering lead to different clusterings in terms of number of clusters and alignment. In the analysis hereby presented, the alignment of clusters was evaluated both at the profile and lesion level. Further analysis should be performed on different and larger cohorts to validate the procedure. The GDM was thoroughly selected based on formal arguments about ordinal data and the incommensurability of NIHSS items. A limitation of this approach is its uncommon use, and alternatives such as cosine or Hamming distances may give different results. This work increases knowledge on patients clustering, a framework that has been less explored in the field, compared to PCA, Factor Analysis, and similar techniques that can be framed as grouping the NIHSS items. These frameworks are not exclusive but dual, in advancing our understanding. The range of application of the GDM and network model clustering, with their statistical features, allow exploitation and validation of our workflow on diverse and multimodal data sets from various domains, from imaging to psychometric and physiological records. For these reasons, unsupervised learning approaches are promising for creating patient-centered models of diseases in the context of AI applications for precision medicine in stroke [19]. Unsupervised characterizations of data, e.g. phenotypization of diseases, could be used to supervise machine learning models on more specific tasks (weak supervision) as well as to provide meaningful insight for

self-supervised learning. Overall the presented methods are general, open source and portable to other problems and disciplines, paving the way for related applications in other biomedical domains.

5 Conclusions

The complexity of the medical domain, including the high number of variables involved, their statistical properties and variety, make it difficult to create data driven mathematical models of diseases from healthcare data. The development of such methods, however, is improving. In this paper we present a paradigm to analyze post-stroke syndromes, as a workflow with two branches. First, we compute distances between NIHSS profiles, scored in acute phase, by means of the General Distance Measure for ordinal data. After turning GDM into GSM, and building a similarity network of patients, the workflow branches in Repeated Spectral Clustering and Affinity Propagation. After each clustering, we compute density maps for the frequency of brain lesions, as segmented by an expert, and stratified by cluster. These maps show separation with one another, even though they are based on behavioral data only, and are aligned with recorded and expected localization of deficits. The clusters obtained with the two different methods validate each other as there is a strong correspondence between them. They describe a fine-grained structure of neurological syndromes, providing additional topographical and etiological insights for the clinician in the hyperacute phase of stroke. Finally, they confirm known limitations of a widely used clinical scale. Challenges related to the application of machine learning, unsupervised learning in particular, still exist, especially regarding the inherent differences in aspects of data captured by different algorithms. However approaches such as clustering demonstrate, as highlighted by this work, a high potential for modeling complex diseases and syndromes in the biomedical domain, especially when exploring hypotheses and effects of different assumptions on large and varied datasets. The presented methods are general, open source and portable to other datasets and disciplines, paving the way for related applications in other biomedical domains.

Ethics statement

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.

Acknowledgements

The principal investigators wish to thank the patients who participated in the study for their time and effort.

Sources of 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.

Disclosures

None.

References

[1] Kathleen Strong, Colin Mathers, and Ruth Bonita. "Preventing stroke: saving lives around the world". In: *The Lancet Neurology* 6.2 (2007), pp. 182–187.

[2] Ali Zandieh et al. "The Underlying Factor Structure of National Institutes of Health Stroke Scale: An Exploratory Factor Analysis". en. In: (2012), p. 6.

[3] Patrick Lyden et al. "Factor Analysis of the National Institutes of Health Stroke Scale in Patients With Large Strokes". In: *Archives of Neurology* 61.11 (Nov. 2004), pp. 1677–1680. ISSN: 0003-9942. DOI: 10.1001/archneur.61.11.1677. URL: https://doi.org/10.1001/archneur.61.11.1677 (visited on 08/30/2022).

[4] Maurizio Corbetta et al. "Common behavioral clusters and subcortical anatomy in stroke". In: *Neuron* 85.5 (Mar. 2015), pp. 927–941. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2015.02.027. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4646844/ (visited on 08/30/2022).

[5] Antonio Luigi Bisogno et al. "A low-dimensional structure of neurological impairment in stroke". In: *Brain Communications* 3.2 (2021), fcab119.

[6] K. Jajuga, M. Walesiak, and A. Bak. "On The General Distance Measure". en. In: *Exploratory Data Analysis in Empirical Research*. Ed. by H.-H. Bock et al. Series Title: Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 104–109. ISBN: 978-3-540-44183-0 978-3-642-55721-7. DOI: 10.1007/978-3-642-55721-7_12. URL: http://link.springer.com/10.1007/978-3-642-55721-7_12 (visited on 08/31/2022).

[7] Brendan J. Frey and Delbert Dueck. "Clustering by Passing Messages Between Data Points". en. In: *Science* 315.5814 (Feb. 2007), pp. 972–976. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1136800. URL: https://www.science.org/doi/10.1126/science.1136800 (visited on 12/19/2022).

[8] A.L.N. Fred and A.K. Jain. "Data clustering using evidence accumulation". In: *2002 International Conference on Pattern Recognition*. Vol. 4. ISSN: 1051-4651. Aug. 2002, 276–280 vol.4. DOI: 10.1109/ICPR.2002.1047450.

[9] Vladimir Fonov et al. "Unbiased average age-appropriate atlases for pediatric studies". In: *NeuroImage* 54.1 (2011), pp. 313–327.

[10] Ulrike von Luxburg. "A tutorial on spectral clustering". en. In: *Statistics and Computing* 17.4 (Dec. 2007), pp. 395–416. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-007-9033-z. URL: http://link.springer.com/10.1007/s11222-007-9033-z (visited on 06/22/2022).

[11] Mariá Cristina Vasconcelos Nascimento, Franklina Maria Bragion de Toledo, and André C. Ponce Leon Ferreira Carvalho. "Consensus Clustering Using Spectral Theory". en. In: *Advances in Neuro-Information Processing*. Ed. by Mario Köppen, Nikola Kasabov, and George Coghill. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 461–468. ISBN: 978-3-642-02490-0. DOI: 10.1007/978-3-642-02490-0_57.

[12] Thomas M. J. Fruchterman and Edward M. Reingold. *Graph Drawing by Force-directed Placement*. 1991.

[13] Robert W. Regenhardt et al. "Association of Infarct Topography and Outcome After Endovascular Thrombectomy in Patients With Acute Ischemic Stroke". en. In: *Neurology* 98.11 (Mar. 2022), e1094–e1103. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/WNL.0000000000200034. URL: http://www.neurology.org/lookup/doi/10.1212/WNL.0000000000200034 (visited on 10/04/2022).

[14] Fana Alemseged et al. "Posterior National Institutes of Health Stroke Scale improves prognostic accuracy in posterior circulation stroke". In: *Stroke* 53.4 (2022), pp. 1247–1255.

[15] Emilia Vitti et al. "Left hemisphere bias of NIH Stroke Scale is most severe for middle cerebral artery strokes". In: *Frontiers in Neurology* 13 (2022), p. 912782.

[16] Brett C Meyer et al. "Modified National Institutes of Health Stroke Scale for use in stroke clinical trials: prospective reliability and validity". In: *Stroke* 33.5 (2002), pp. 1261–1266.

[17] Brett C Meyer and Patrick D Lyden. "The modified National Institutes of Health Stroke Scale: its time has come". In: *International journal of stroke* 4.4 (2009), pp. 267–273.

[18] Antonio Luigi Bisogno et al. "The Oxford cognitive screen (OCS) as an acute predictor of long-term functional outcome in a prospective sample of stroke patients". In: *Cortex* 166 (2023), pp. 33–42.

[19] Anna Bonkhoff and Christian Grefkes. "Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence". In: *Brain : a journal of neurology* 145 (Apr. 2022), pp. 457–475. DOI: 10.1093/brain/awab439.

[20] Andrew Ng, Michael Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm". In: *Advances in neural information processing systems* 14 (2001).

[21] Alexander Strehl and Joydeep Ghosh. "Cluster ensembles—a knowledge reuse framework for combining multiple partitions". In: *Journal of machine learning research* 3.Dec (2002), pp. 583–617.

[22] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. "Factor graphs and the sum-product algorithm". In: *IEEE Transactions on information theory* 47.2 (2001), pp. 498–519.

Tables

Repeated Spectral Clustering			Affinity Propagation		
	Deficit (NIHSS item)	Lesion Location		Deficit (NIHSS item)	Lesion Location
C0	Motor Arm Left, Motor Leg Left, Inattention, Facial Palsy, Dysarthria, Sensory, Visual	Right cortical and subcortical frontal and parietal lobes, right basal ganglia and internal capsule	C0	LOC-Q, Best Language, Dysarthria, Facial Palsy, Visual, Best Gaze	Left cortico-subcortical frontal areas
C1	Visual, Best Gaze, Inattention, Limb Ataxia, Sensory	Right cortico-subcortical frontal lobe, bilateral cortico-subcortical parietal and occipital lobes	C1	Motor Arm Left, Motor Leg Left, Visual, Inattention, Facial Palsy, Dysarthria, Sensory	Right cortical and subcortical frontal and parietal lobes, right basal ganglia and internal capsule
C2	Motor Arm Right, Motor Leg Right, LOC-Q, Best Language, Facial Palsy, Dysarthria, Limb Ataxia	Left subcortical frontal lobe, left basal ganglia and internal capsule	C2	Motor Arm Left, Motor Arm Right, Limb Ataxia, Facial Palsy, Best Language, Visual	Bilateral subcortical frontal lobe, bilateral occipital and parietal lobe
C3	Motor Arm L, Motor arm R, Sensory, Facial Palsy, Dysarthria, Visual, Inattention	Bilateral basal ganglia and internal capsule	C3	Motor Arm Right, Motor Leg Right, Dysarthria, Facial Palsy, Best Language, LOC-Q	Left subcortical frontal areas, left basal ganglia and internal capsule
C4	LOC-Q, LOC-C, Best Language, Dysarthria, Visual, Best Gaze, Facial Palsy, Sensory	Left cortical and subcortical frontal and parietal lobes	C4	LOC-Q, Best language, Visual	Left cortical frontal and parietal areas
			C5	Visual, Facial Palsy, Inattention	Right insula, cortical frontal and parietal areas
			C6	Motor Arm Left, Best Gaze, Visual, Limb Ataxia, Sensory	Bilateral brainstem, bilateral cerebellum, bilateral temporal and occipital areas and thalami

Table 1: Table of cluster characteristic items, and lesion locations evidenced in brain scans.

Figures with Figures Legends

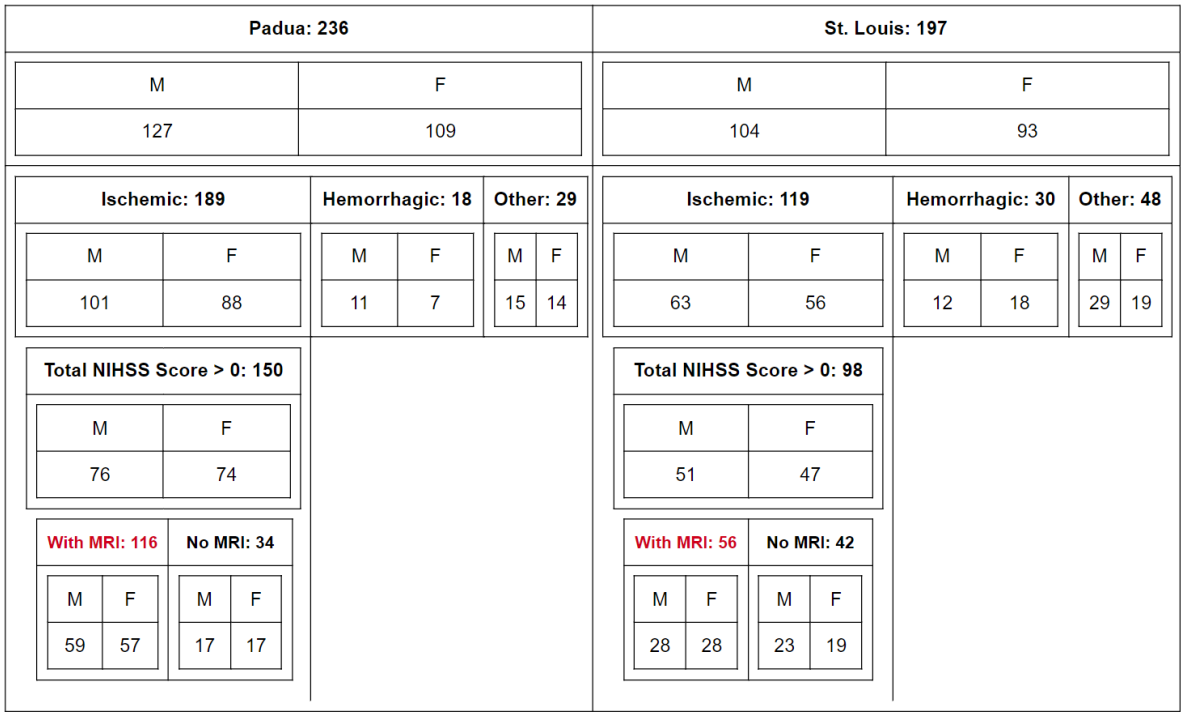

Figure 1: Patient selection workflow. Red marks patients taken into account for clustering. Abbreviations: M: male, F: female.

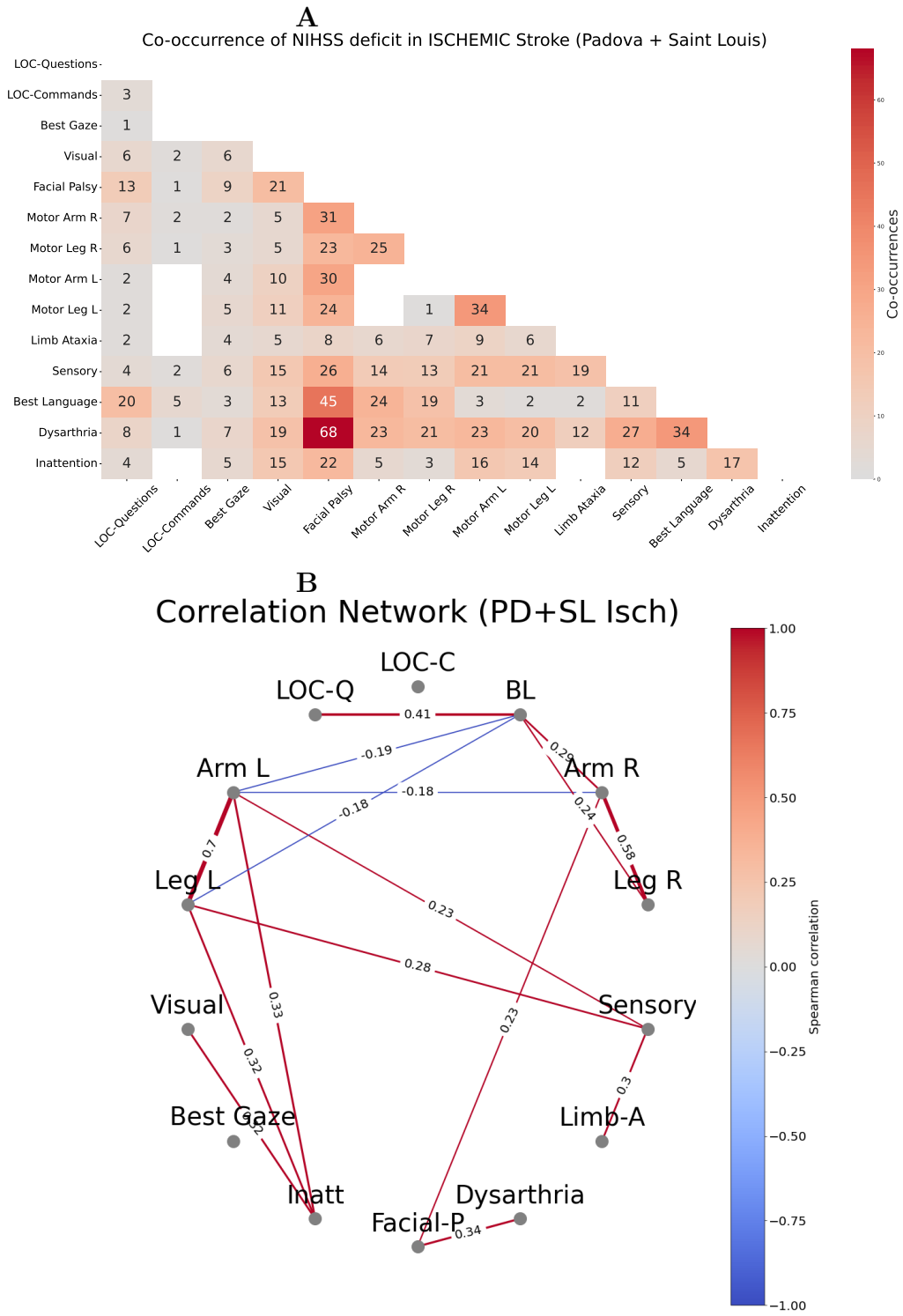

Figure 2: (A) Table visualization of co-occurrences of item deficits. Intense red for higher number of co-occurrences. Empty blocks represent zero co-occurrences. (B) Graph visualization of correlation of item deficits severity. Only significant correlations ($\alpha < 0.05$, Benjamini-Hochberg correction) visualized. PD+SL Isch for Padua and Saint Louis Ischemic data. 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.

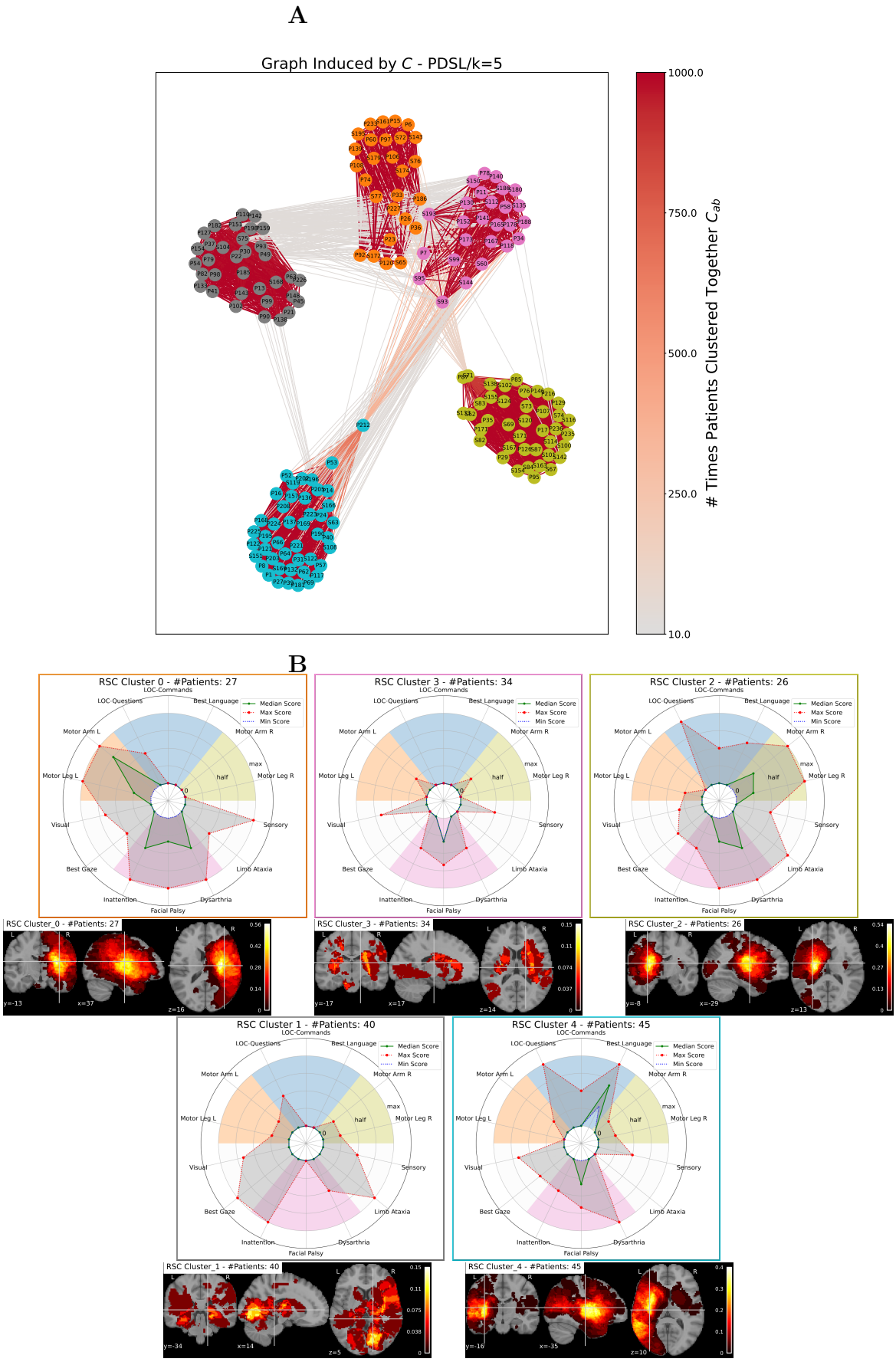

Figure 3: (A) Graph visualization of the co-occurrence matrix C. Darker red color links patients frequently clustered together. (B) Radar charts and lesion heat maps for the clusters obtained from RSC. PDSL stands for Padua and St. Louis data.

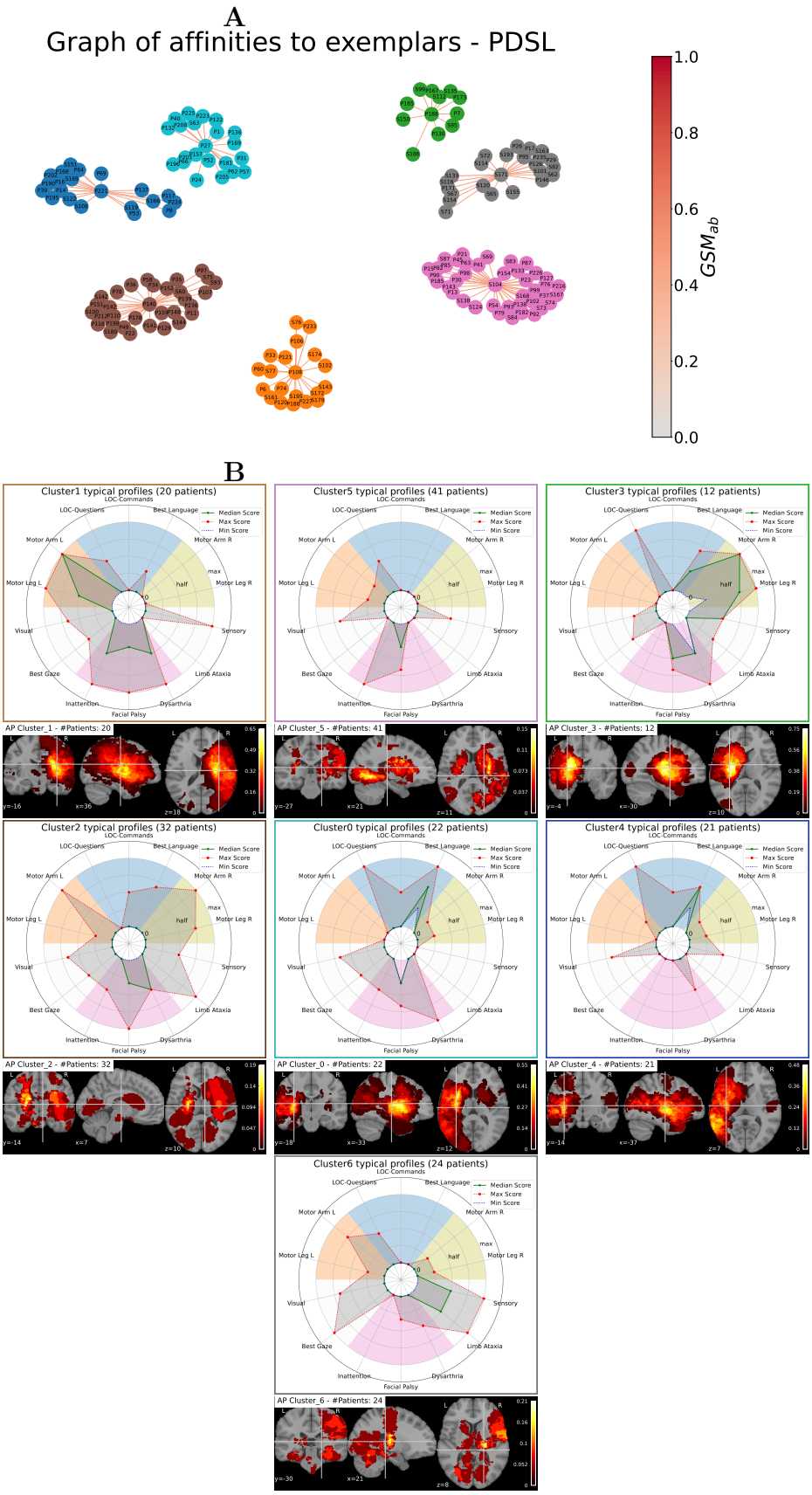

Figure 4: (A) Graph visualization of the element-to-exemplar affinity. Darker red color means linked patients are more similar. (B) Radar charts and lesion heat maps for the clusters obtained from Affinity Propagation. PDSL stands for Padua and St. Louis data.

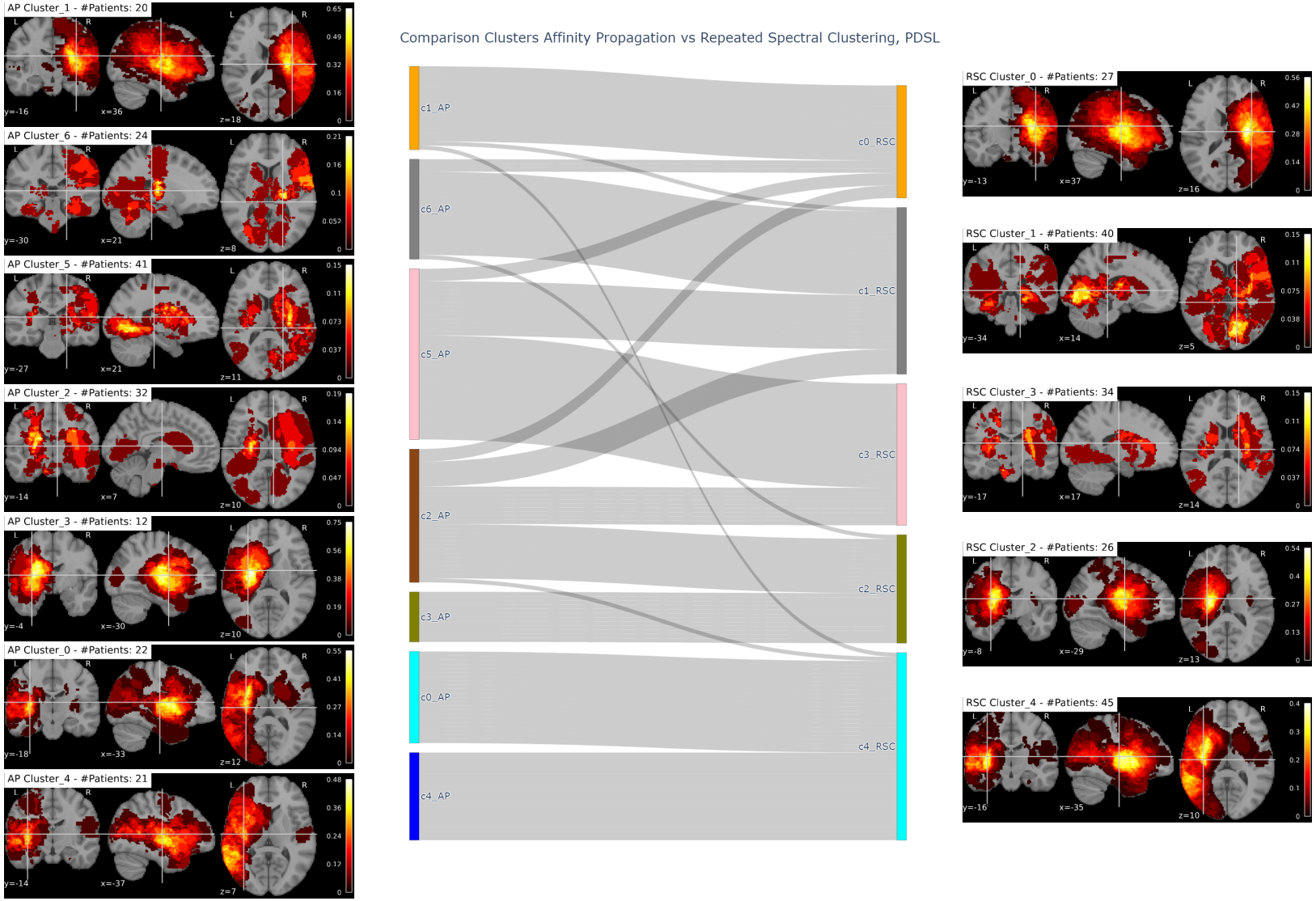

Figure 5: Sankey plot showing a direct comparison between affinity propagation (AP) and repeated spectral clustering (RSC) results. Patients with similar symptoms have brain lesions that overlap within their clusters. PDSL stays for Padua and St. Louis.

Supplemental Materials

Supplemental Methods

Figures S1 - S8

References [20] [21] [8] [11] [22] [7]