Revealing heterogeneity in dementia using data-driven unsupervised clustering of cognitive profiles

[1,2]Sayantan Kumar, MS[§], [2]Inez Y. Oh, PhD, [3]Suzanne E. Schindler, PhD, [3,4]Nupur Ghoshal, MD, [2]Zachary Abrams[†], PhD, [1,2]Philip Payne, PhD

[1] Department of Computer Science and Engineering, McKelvey School of Engineering, Washington University in St Louis, St. Louis, Missouri, USA

[2] Institute for Informatics, Data Science and Biostatistics (I2DB), Washington University School of Medicine, St. Louis, Missouri, USA

[3] Division of Neurology, Washington University School of Medicine, St Louis, USA

[4] Department of Psychiatry, Washington University School of Medicine, St Louis, USA

§Corresponding author:

Sayantan Kumar, MS
660 South Euclid Ave, Campus Box 8132
St Louis, Missouri 63110
sayantan.kumar@wustl.edu
(+1)314.326.9095

Keywords: Dementia, heterogeneity, Clinical Dementia Rating, SillyPutty, hierarchical clustering, subtypes, cognitive profiles, disease progression

Abstract

Dementia is characterized by a decline in memory and thinking that is significant enough to impair function in activities of daily living. Patients seen in dementia specialty clinics are highly heterogenous with a variety of different symptoms that progress at different rates. Recent research has focused on finding data-driven subtypes for revealing new insights into dementia's underlying heterogeneity, compared to analyzing the entire cohort as a single homogeneous group. However, existing studies on dementia subtyping suffer from the following limitations: (i) focusing on AD-related dementia only and not examining heterogeneity within dementia as a whole, (ii) using only cross-sectional baseline visit information for clustering and (iii) predominantly relying on expensive imaging biomarkers as features for clustering. In this study, we used a data-driven unsupervised clustering algorithm named SillyPutty, in combination with hierarchical clustering on neuropsychological assessment scores to estimate subtypes within a real-world clinical dementia cohort. We incorporated all longitudinal patient visits into our clustering analysis, instead of relying only on baseline visits, allowing us to explore the ongoing relationship between subtypes and disease progression over time. Results showed evidence of (i) subtypes with very mild or mild dementia being more heterogenous in their cognitive profiles and risk of disease progression.

1. Introduction

Dementia is characterized by a notable deterioration in memory and cognitive function, that is significant enough to impair function in activities of daily living. It stems from various factors, including reversible causes like medication-induced cognitive impairment, and irreversible ones such as progressive neurodegenerative conditions[1–3]. Alzheimer's Disease (AD) is the predominant cause of dementia in older adults, yet various other conditions like cerebrovascular disease and disorders linked to Lewy bodies, tau tangles, or limbic-predominant age-related TDP-43 encephalopathy (LATE)[4]. Dementia patients exhibit diverse symptoms and progress at varying rates, likely influenced by underlying brain pathologies, baseline cognitive ability, genetic factors, medical conditions, and social determinants of health. Consequently, patients treated in dementia clinics manifest significant heterogeneity, representing a spectrum of dementia types[5–7]. A more thorough understanding of this clinical heterogeneity could enhance dementia diagnosis and prognosis, facilitating tailored care for patients and their families[8,9].

Previous studies have often relied on predefined neuropathological categories to identify subtypes within clinical dementia populations[10–12]. However, these approaches typically categorize pathology subtypes based on subjective clinical diagnoses rather than objective, data-driven methods[13,14]. Recently, the widespread availability of Electronic Health Records (EHR) data alongside advanced machine learning techniques has enabled data-driven approaches for revealing new insights into dementia's underlying heterogeneity[15]. For instance, clustering algorithms can stratify dementia patients into subtypes based on key EHR features, enhancing predictive ability compared to analyzing the entire cohort as a single homogeneous group[16–18]. However, current research on dementia subtyping faces three main limitations. First, there is a significant focus on parsing

heterogeneity within AD-related dementia potentially neglecting important insights that could arise from examining heterogeneity within dementia as a whole[19]. Second, conventional clustering methods used in dementia research have analysed cross-sectional single time-point data; however, given the heterogeneity in disease progression, mapping longitudinal trajectories is an important focus for dementia research[20–22]. Finally, existing research predominantly relies on imaging biomarkers obtained through expensive procedures, such as magnetic resonance imaging (MRI) and positron emission tomography (PET) scans. However, existing research often overlooks the potential of using non-imaging clinical data, such as routinely collected electronic health records (EHR), as a valuable, cost-effective, and non-invasive solution for addressing heterogeneity in dementia[15,23,24]. Addressing these gaps is essential for effective clinical decision-making and precision diagnostics tailored to each subtype.

Our goal was to analyze the heterogeneity of cognitive performance and disease progression within a real-world clinical dementia cohort. Towards this end, we aimed to delineate subtypes using a newly-developed unsupervised clustering technique called SillyPutty on neuropsychological assessment scores of patients seen in a memory clinic (Figure 1). This new heuristic clustering approach started with a set of clustering assignments obtained from hierarchical clustering and iteratively adjusted cluster assignments to maximize the average silhouette width[25]. A novel aspect of our approach is the inclusion of all longitudinal patient visits for clustering analysis, rather than solely relying on baseline visits, which allowed us to examine the longitudinal relationship between subtypes and disease progression. Specifically, we aimed to (i) examine how subtypes at similar stages of dementia are heterogeneous in their cognitive characteristics (ii) analyze patient transitions between different subtypes across multiple visits to examine the heterogeneity (variability) in progression rate amongst patients at different stages of dementia (Figure 1).

2. Materials and Methods

2.1 Data sources and study participants

This retrospective study analyzed electronic health records (EHR) data extracted from the Washington University in St. Louis Research Data Core (RDC), which serves as a repository for patient clinical data sourced from BJC HealthCare and Washington University Physicians. Approval for this study was obtained from the Washington University Institutional Review Board, which granted a waiver of HIPAA Authorization for the use of Protected Health Information (PHI). The study cohort included all dementia patients treated at the Memory Diagnostic Center (MDC) at Washington University School of Medicine for evaluation of memory and/or thinking concerns (Figure 1). The dataset, sourced from Allscripts TouchWorks, included office visits with neuropsychological assessment measures recorded between June 1, 2012, and May 31, 2018. This timeframe was chosen to mitigate potential data harmonization challenges arising from an EHR system transition that commenced on June 1, 2018.

The Washington University MDC required patients to have insurance coverage and accompany a support person familiar with their history to appointments. Patients underwent thorough assessments, including comprehensive history-taking and neurologic examinations. Memory specialists utilized the Clinical Dementia Rating (CDR®) scale to evaluate dementia severity, where scores ranged from 0 (normal cognition) to 3 (severe dementia). The CDR assessed changes in memory, thinking, and function compared to previous abilities[26]. Specialists documented CDR scores in patients' records, which were extracted from the structured EHR data. Each patient visit with a recorded CDR score at that particular visit were eligible for inclusion in our study (Figure 1).

2.2 Feature selection and preprocessing

Our initial set of feature variables consisted of cognitive assessment scores, indicating performance on a standardized battery of cognitive tasks to assess the severity of dementia. Compared to expensive and/or invasive procedures like neuroimaging biomarkers, these scores are standard metrics in dementia research and is recorded for all patients in the memory clinic. We performed feature selection to determine the optimal set of features for clustering. We started with two sets of features as follows: (i) Cognitive assessment scores: Boston Naming Test[27], Mini-Mental State Exam[28], Short Blessed[29], Word List Memory Task[30], Verbal Fluency[31] and (ii) six components of CDR score: Memory, Orientation, Judgment and Problem Solving, Community Affairs, Home and Hobbies, and Personal Care. The CDR components assess six cognitive and functional domains, each of which provides critical insights into different aspects of a patient's cognitive health respectively. The component scores range between {0,0.5,1,2,3} with 0 indicating normal cognition and 3 indicating severe impairment. The CDR components assess performance of dementia patients in multiple cognitive and functional domains, providing critical insights into different aspects of a patient's cognitive health. More detailed information about each component scores are available in the Supplementary material.

We plotted the *t*-distributed stochastic neighbour embedding (T-SNE) distribution for the 2 feature sets to determine the optimal set of features for clustering (Figure 1)[32]. For both the feature sets, the correlation matrix heatmap between the individual variables are presented in the Supplementary material (eFigure 1). Feature pre-processing included clipping the outlier values to the 5th and 95th percentile values and scaling between [0,1] using the Minmax Scalar package from sklearn. Since each of the selected features are ordinal variables and had a very low missing rate of <5%, we followed the strategy in [33]

and imputed the missing values for each feature column using the median value of that feature across all visits (Figure 1).

2.3 Unsupervised clustering algorithm

In our analyses, we used the SillyPutty algorithm [25] for clustering. SillyPutty is a heuristic algorithm that utilizes the concept of silhouette widths to optimize cluster assignments. The silhouette width, widely recognized and utilized, measures how well each data point fits its designated cluster. Ranging from -1 to 1, a value close to 1 indicates strong clustering, with clear distinctions between clusters and high cohesion within each group. Conversely, a value nearing 0 suggests overlapping clusters, while a value close to -1 indicates potential misclassification of the data point.

The objective of SillyPutty is to iteratively adjust cluster assignments to maximize the average silhouette width. Initially, SillyPutty starts with any set of cluster assignments, which can be user-specified, randomly chosen, or obtained from other clustering techniques. The algorithm then enters a loop where it iteratively refines the clustering. At each iteration, SillyPutty calculates silhouette widths for the current clustering. It identifies the data point with the lowest (most negative) silhouette width, indicating a potential misclassification, and reassigns this point to the cluster it is closest to. This process continues until all data points have non-negative silhouette widths, showing convergence or until either of the following early termination condition are met : (i) the maximum number of iterations is reached to prevent infinite loops and (ii) the same silhouette width vector is detected within the last n iterations (user-specified), indicating a potential infinite loop. Upon completion, SillyPutty returns the refined clustering, silhouette widths, and other relevant information.The pseudo code for the SillyPutty algorithm is given in the Supplementary Material.

The standalone SillyPutty algorithm starts with purely random cluster assignments, repeating the algorithm with different random starting points (Random SillyPutty). Since

SillyPutty can start with any cluster assignments, we initialized the cluster assignments with hierarchical clustering before applying the SillyPutty algorithm (Figure 1). This choice is motivated by results in the original SillyPutty paper where hierarchical SillyPutty outperformed other clustering methods such as Random SillyPutty, Partition Against Medoids (PAM) and hierarchical clustering on multiple simulated datasets. The optimal number of clusters for SillyPutty was estimated by using the Hierarchical SillyPutty algorithm to determine the best mean silhouette width, for a range of clusters values (K) from K = 2 to 16.

2.4 Clustering on longitudinal visits

In our study, clustering analyses were performed on all data that was available from a 6-year period between 2012-2018, and some patients had multiple visits during this period. For the clustering step, each visit was assumed independent without any temporality (linkage) between individual visits of the same patient. The temporality information between individual visits was used downstream for analyzing the relationship between different clusters. This approach – including all visits for clustering as opposed to including the baseline visit of each patient was taken to enable downstream longitudinal analysis and track the symptom progression rate among patients. For example, at any given point in time, a patient exists in a single cluster (dementia subtype), but transitions between different clusters over time (Figure 1).

2.5 Clustering evaluation

SillyPutty with hierarchical clustering was compared with the following state-of-the-art clustering techniques: (i) Random SillyPutty (SillyPutty with random initial assignments), (ii) hierarchical clustering, and (iii) Partition Against Medoids (PAM). All clustering algorithms were qualitatively evaluated using the metrics used in the original SillyPutty paper, such as

Mean Silhouette Widths (MSW)[25,34], T-SNE[32] and multidimensional scaling (MDS)[35].

We defined the cognitive and comorbidity profile of each dementia subtype and examined both the inter-subtype and intra-subtype variability To gain an insight into which patients have a higher probability of progression, patient transitions between different dementia subtypes were analysed across multiple visits. Patients with only a single visit were censored from our analysis. Finally, the differences in progression rate, both within and between CDR score categories were measured.

2.6 Software packages and code availability

The SillyPutty algorithm was implemented using the R package published in the Comprehensive R Archive Network (CRAN) https://cran.r-project.org/web/packages/SillyPutty/index.html. All other visualizations were performed using the seaborn package and Python 3.7. The implementation code for this project will be made available in GitHub upon acceptance.

3. Results

3.1 Sample characteristics

Longitudinal data from 1,845 patients with 2,737 visits were eligible for inclusion, where each visit recorded a CDR score. 953 patients have a single visit and the remaining 892 patients have multiple visits (maximum number of visits = 5) with a visit interval of (mean +/- STD) 6.8 +/- 3.6 months. The median age of the cohort at first encounter of was 79 years with 61% of the patients being female. In terms of race and ethnicity, 83% patients were White and 14% patients were Black or African American; 97% patients were Non-Hispanic or Latino, while 1.3% were Hispanic or Latino (Table 1). The most frequently occurring brain-related disorders in our cohort included memory loss (N = 987 patients; 54 %),

Alzheimer's Disease (N = 1061 patients; 57.5%), Parkinsonism (N = 66 patients; 3.6%), Major Depressive Disorder (N = 455 patients; 24.7%) and Obstructive Sleep Apnea (N = 366 patients; 19.8%) respectively (Table 1).

3.2 Optimal set of features for clustering

We can see that using the cognitive scores as features creates a gradient from low to high CDR but did not result in clusters (Figure 2A). However, using the CDR components as features created clusters that were distinct without any significant overlap across the CDR categories (Figure 2B). Analysing the degree of overlap of points across the different CDR categories, the six components of CDR were selected as the optimal set of features for clustering. Another motivation behind selecting the individual CDR component as features compared to the aggregate CDR score is the fact that for patients having the same CDR score, the individual components might be different from one another, allowing us a more granular approach of studying the sub-phenotyping of patients. The correlation between different feature variables for both CDR components and cognitive scores are shown in the Supplementary material (eFigure 1) respectively.

3.3 Clustering evaluation

3.3.1 Optimal number of clusters

Comparing the mean silhouette width (MSW) values of different values of the number of clusters, K = 4, 6, 10 and 15 were identified as potential candidates to be the optimal number of clusters (Figure 3). The rationale behind the selection of these particular configurations of K was that the MSW values for these K were higher compared to the neighbouring K points. This observation is line with the T-SNE distribution across different clusters for each of K = 4, 6, 10 and 15 (Supplemental material; eFigure 2) and the T-SNE distribution across CDR stages (Figure 2B). The visits in the earlier stages of dementia (e.g. CDR = 0.5) can

potentially form multiple clusters indicating a level of heterogeneity in cognitive characteristics can be better captured for higher number of clusters (K = 10 and K = 15) compared to lower values (K = 4 and K = 6) (eFigure 2). The T-SNE plots for K=10 and K=15 were identical for early stages of dementia (CDR = 0,0.5) with smaller subclusters for CDR >= 1. In our subsequent analyses, we selected K = 10 as the optimal number of clusters.

3.3.2 Comparison with other clustering methods

SillyPutty initialized with cluster assignments generated by hierarchical clustering showed higher mean silhouette width (MSW = 0.35715) across clusters compared to SillyPutty initialized with random cluster assignments (Random SillyPutty; MSW = 0.19989) and other state-of-the-art clustering algorithms like PAM (MSW = 0.3391) and hierarchical clustering (MSW = 0.30936) (Figure 4). Random SillyPutty performed the worst with 9 out of 10 clusters having negative silhouette widths indicating the possibility of misclassification of points to the wrong cluster or overlap between the clusters. Using only hierarchical clustering demonstrated negative silhouette widths in 7 out of 10 clusters. However, using the SillyPutty algorithm in combination with hierarchical clustering (HSP) showed positive silhouette widths for all the 10 clusters (Figure 4). The T-SNE and MDS plots for hierarchical SillyPutty showed more distinct and well-separated clusters compared to Random SillyPutty, PAM and hierarchical clustering (Figure 4).

3.4 Subtype demographics

Subtypes (clusters) can either be homogenous, having a unique CDR (e.g. C_9, C_4), or composite, including two different CDR scores, (e.g. C_7, C_5) (Figure 5). C_2 was a predominantly healthy cluster with patient visits associated with no cognitive impairment. C_4 and C_9 only consisted of visits with very mild dementia, while C_7, C_5, C_8 and C_6 are mostly dominated by visits with mild dementia. Patient visits at a more advanced stages of the

disease were mostly concentrated in clusters C_{10}, $C_{3,}$ and C_1 respectively. Higher number of subtypes were associated with the mild dementia (CDR <= 1) compared to moderate to severe dementia (CDR > 1). This is in line with the T-SNE distributions in Figure 2B and eFigure 2, where data points with CDR = 0.5 and 1 are more scattered, potentially forming multiple subtypes.

Demographic characteristics of patients within each subtype are presented in Table 2. Patents with no cognitive impairment (C_2) were generally younger than the patients in other subtypes with mild to severe dementia. Patients across all 10 subtypes were mostly white and non-Hispanic (Table 2). Subtypes with very mild or mild dementia ($C_{4,}$ $C_{9,}$ C_7, $C_{5,}$ C_8 and C_6) included a higher proportion of patients with memory issues, depression and sleep disorders compared to the subtypes with moderate or severe dementia (Table 2). Dementia patients diagnosed with AD were mostly assigned to the subtypes in the later stages of dementia (C_{10}, $C_{3,}$ and C_1). For all the six cognitive assessment scores: Boston Naming Test, Mini-Mental State Exam, Short Blessed, Word List Memory Task, Logical Memory and Verbal Fluency, we observed a natural progression of increasing cognitive impairment from early (mild) to later (moderate to severe) dementia stages, but not much variability across subtypes with same CDR score (e.g. C_4 and C_9) (Table 2).

3.5 Cognitive characteristics of identified subtypes

The association between the six CDR components and dementia subtypes allows for interpreting the cognitive profiles of these subtypes (Figure 6). The significance of these component scores lies in their ability to provide a detailed picture of the multifaceted nature of cognitive decline in dementia. For all the six CDR components, there is a natural progression of increasing cognitive impairment from early (mild) to later (moderate to severe) dementia stages. In our analyses, we analyzed the inter-subtype variability to examine if subtypes with the same CDR are heterogeneous in their CDR component scores (Figure 6).

Further we also defined the cognitive profile of each subtype based on the distribution of each component score of patients within that subtype (Table 3).

Subtype C_2 consisted of a relatively healthy population with minimal cognitive and functional impairments (Table 3). Patients in both C_4 and C_9 recorded the same CDR score of 0.5 and exhibited similar trends of mild to moderate memory loss and fully capable of self-care (Figure 6). However, C_4 visits were associated with slight disorientation, and slight impairment in judgement, community and home activities compared to patients in C_9 who were fully functional and healthy in those domains (Table 3). Comparing the subtypes (C_7, C_5, C_8 and C_6) with predominantly visits characterized by mild dementia (CDR = 1), we observed that the four subtypes showed similar level of moderate memory loss and orientation problems. However, patient visits in C_8 and C_6 showed more severe impairment in judgement & problem-solving ability, along with home and community activities compared to visits in C_7 and C_5 (Figure 6, Table 3). We observed considerable variability in self-care abilities between C_7, C_5, C_8 and C_6 respectively. While C_7 patients were fully capable of self-care, those in C_5, C_8 and C_6 required assistance in daily personal activities. The subtypes (C_{10}, C_3, and C_1) with predominantly visits characterized by moderate to severe dementia (CDR >1) showed similar extent of severe cognitive impairments in all the domains except personal care. While C_{10} patients only needed prompting for daily activities, patients in C_3, and C_1 required significant assistance in dressing, hygiene and keeping of personal effects.

Overall, it can be concluded that subtypes at the early stages of dementia demonstrated more heterogeneity in cognitive characteristics compared to the ones at the later stages of the disease. The underlying heterogeneity within the CDR component scores justifies our choice of selecting the 6 CDR component measures as our optimal set of features for clustering instead of other cognitive assessment scores.

3.6 Cluster transitions

Finally, we examined how patients transitioned between the different subtypes over time and how these transitions were related to disease progression from CDR <= 0.5 to CDR = 1 (Figure 7A) and from CDR = 1 to CDR >1 (Figure 7B). With clustering performed on all visits of all patients in our cohort as opposed to only the baseline visit, we assumed that at any given point in time, a patient exists in a single subtype but transitions to a different subtype in their next visit. Our goal was to assess if subtypes at similar stages of dementia have variability in the risk of progression of advanced dementia stages. For example, patient visits in both C_9 and C_4 were characterized by very mild dementia (CDR = 0.5), but patients in C_4 had more transitions to visits with mild dementia (CDR = 1; C_5 C_6 and C_7), indicating greater chances of disease progression for C_4 visits (Figure 7A). On the contrary, majority of the transitions from C_9 were either within the same subtype (self-transitions) or to subtype C_4 at similar disease stages (CDR = 0.5). This indicates the variability in the risk of progression for patient visits in 2 different subtypes at similar stages of dementia. Analyzing the subtype transitions from mild dementia (CDR = 1) to moderate/severe dementia (CDR > 1), we observed variability in the chances of progression amongst subtypes with predominantly CDR = 1 visits (C_5, C_6, C_7 and C_8). C_5 and C_7 have more transitions to subtypes (C_1, C_3 and C_{10}) with moderate to severe dementia (CDR > 1) compared to C_6 and C_8 (Figure 7B). The number of transitions between all every pair of 10 subtypes are presented in eTable 1.

4. Discussion

In this study, we used the state-of-the-art unsupervised clustering method named SillyPutty, in combination with hierarchical clustering to estimate subtypes within a real-world clinical dementia cohort. We incorporated all longitudinal patient visits into our clustering analysis, instead of relying only on baseline visits, allowing us to explore the ongoing relationship

between subtypes and disease progression over time. Results showed evidence of (i) subtypes with very mild or mild dementia being more heterogenous in their cognitive profiles and risk of disease progression.

Neuroimaging biomarkers such as MRI and PET have been predominantly used for estimating subtypes to assess heterogeneity in AD. Compared to MRI, which requires specialized equipment and facilities, neuropsychological assessment scores are more easily accessible and can be conducted in various settings, such as clinics and homes, making them suitable for large-scale screenings. The six component measures of CDR used in our analyses examined different cognitive and functional domains providing critical insights into different aspects of a patient's cognitive health due to dementia. These component measures are closely linked to functional abilities and daily living skills, which are critical for evaluating the impact of dementia on a patient's quality of life. Further, our results demonstrate that patients with the same CDR can have different cognitive characteristics with respect to the different CDR components. We believe that using the CDR component scores as clustering features is a cost-effective promising approach to investigate cognitive heterogeneity in dementia, which can be helpful effective clinical decision-making and precision diagnostics tailored to each subtype.

Results showed that both subtypes C_4 and C_9 were at early stages of dementia with mild cognitive impairment (MCI; CDR = 0.5), but had different cognitive profiles and risk of progression. For example, higher number of transitions to advanced CDR stages for C_4 compared to C_9 indicated that patients in C_4 and C_9 can be considered as progressive MCI and stable MCI respectively. Further, C_4 patients have more impairment in orientation, judgement & problem solving, home & hobbies and community affairs compared to patients in C_9. These particular cognitive and functional domains might be useful features to distinguish between patients with stable or progressive cognitive decline. Our results also indicated that

subtypes with visits having mild dementia were more heterogenous in their cognitive profiles compared to subtypes with visits having moderate to severe dementia. Identifying and diagnosing individuals in the initial stages of their cognitive decline early can enable timely intervention, which might slow the progression of the disease and improve the quality of life.

In this study, the unsupervised clustering analysis was performed using all aggregated visits of every patient, instead of using only their baseline visits. This design choice was motivated by the following reasons. First, analyzing visit-level data enables the examination of temporal patterns and transitions between dementia subtypes over time. This helps in understanding how dementia progresses in different patients and how they may transition between various stages and subtypes, providing insights into the disease's trajectory that are not possible with a baseline-only approach. Second, considering each visit as a separate data point significantly increases the volume of data available for analysis. This is particularly important when the patient cohort is small, as it enhances the statistical power and reliability of the clustering results. Finally, dementia often progresses over many years, but our electronic health record (EHR) data spans only six years (2012-2018). Comprehensive longitudinal analysis is possible only if all visits for each patient are available. However, our dataset included patients with different windows of their clinical care (e.g., 2006-2014, 2016-2022). With the availability of EHR data only from 2012-2018, we chose to cluster at the visit level.

Our study has some important limitations worth considering. First, our pipeline is built on a single EHR dataset. As part of future work, we plan to test the generalizability of our pipeline on additional datasets. Second, the sample size of our dataset is relatively small which restricts our ability to perform rigorous statistical analyses on the variable rate of progression of dementia subtypes. Applying this pipeline to a larger EHR dataset can allow us to estimate probabilistic estimates of disease progression, backed by rigorous statistical

analysis. Finally, our analysis did not adopt a patient-centered perspective, focusing instead on aggregate data rather than individual trajectories. This presents an opportunity for future research to employ a different analytical approach, considering data from a more granular, patient-centric viewpoint.

5. Conclusions

In our study, we applied a recently developed data-driven unsupervised clustering algorithm named hierarchical SillyPutty on the CDR component scores and identified 10 subtypes within a real-world clinical dementia cohort. Unlike previous approaches, we included all longitudinal patient visits in our clustering analysis, which allowed us to investigate the dynamic relationship between subtypes and disease progression over time. Our findings indicated that subtypes with very mild or mild dementia exhibited greater heterogeneity in cognitive profiles and varied risks of disease progression.

References

1. Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, et al. Global prevalence of dementia: a Delphi consensus study. The lancet. 2005;366: 2112–2117.
2. Arvanitakis Z, Shah RC, Bennett DA. Diagnosis and management of dementia. Jama. 2019;322: 1589–1599.
3. Mendez MF, Cummings JL. Dementia: a clinical approach. Butterworth-Heinemann; 2003.
4. Zanetti M, Ballabio C, Abbate C, Cutaia C, Vergani C, Bergamaschini L. Mild cognitive impairment subtypes and vascular dementia in community□dwelling elderly people: a 3□year follow□up study. J Am Geriatr Soc. 2006;54: 580–586.
5. Ryan J, Fransquet P, Wrigglesworth J, Lacaze P. Phenotypic heterogeneity in dementia: a challenge for epidemiology and biomarker studies. Front Public Health. 2018;6: 181.
6. Cohen-Mansfield J. Heterogeneity in dementia: challenges and opportunities. Alzheimer Dis Assoc Disord. 2000;14: 60–63.

7. Mayeux R, Stern Y, Spanton S. Heterogeneity in dementia of the Alzheimer type: evidence of subgroups. Neurology. 1985;35: 453–453.

8. Seelaar H, Rohrer JD, Pijnenburg YAL, Fox NC, Swieten JC van. Clinical, genetic and pathological heterogeneity of frontotemporal dementia: a review. J Neurol Neurosurg Psychiatry. 2011;82: 476–486. doi:10.1136/jnnp.2010.212225

9. Kehagia AA, Barker RA, Robbins TW. Neuropsychological and clinical heterogeneity of cognitive impairment and dementia in patients with Parkinson's disease. Lancet Neurol. 2010;9: 1200–1213. doi:10.1016/S1474-4422(10)70212-X

10. Jack Jr CR, Wiste HJ, Weigand SD, Knopman DS, Mielke MM, Vemuri P, et al. Different definitions of neurodegeneration produce similar amyloid/neurodegeneration biomarker group findings. Brain. 2015;138: 3747–3759.

11. Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. Lancet Neurol. 2011;10: 785–796.

12. Lam B, Masellis M, Freedman M, Stuss DT, Black SE. Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. Alzheimers Res Ther. 2013;5: 1–14.

13. Nettiksimmons J, DeCarli C, Landau S, Beckett L, Initiative ADN. Biological heterogeneity in ADNI amnestic mild cognitive impairment. Alzheimers Dement. 2014;10: 511-521. e1.

14. Noh Y, Jeon S, Lee JM, Seo SW, Kim GH, Cho H, et al. Anatomical heterogeneity of Alzheimer disease: based on cortical thickness on MRIs. Neurology. 2014;83: 1936–1944.

15. Kumar S, Oh I, Schindler S, Lai AM, Payne PR, Gupta A. Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. JAMIA Open. 2021;4: ooab052.

16. Habes M, Grothe MJ, Tunc B, McMillan C, Wolk DA, Davatzikos C. Disentangling heterogeneity in Alzheimer's disease and related dementias using data-driven methods. Biol Psychiatry. 2020;88: 70–82.

17. Varol E, Sotiras A, Davatzikos C, Initiative ADN. HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. Neuroimage. 2017;145: 346–364.

18. Dong A, Honnorat N, Gaonkar B, Davatzikos C. CHIMERA: Clustering of heterogeneous disease effects via distribution matching of imaging patterns. IEEE Trans Med Imaging. 2015;35: 612–621.

19. Turknett J, Wood TR. Demand Coupling Drives Neurodegeneration: A Model of Age-Related Cognitive Decline and Dementia. Cells. 2022;11: 2789. doi:10.3390/cells11182789

20. Young AL, Marinescu RV, Oxtoby NP, Bocchetta M, Yong K, Firth NC, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nat Commun. 2018;9: 4273.

21. Linde RM van der, Dening T, Stephan BCM, Prina AM, Evans E, Brayne C. Longitudinal course of behavioural and psychological symptoms of dementia: systematic review. Br J Psychiatry. 2016;209: 366–377. doi:10.1192/bjp.bp.114.148403

22. Melis RJF, Haaksma ML, Muniz-Terrera G. Understanding and predicting the longitudinal course of dementia. Curr Opin Psychiatry. 2019;32: 123. doi:10.1097/YCO.0000000000000482

23. Alexander N, Alexander DC, Barkhof F, Denaxas S. Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning. BMC Med Inform Decis Mak. 2021;21: 343. doi:10.1186/s12911-021-01693-6

24. Nguyen TTT, Lee H-H, Huang L-K, Hu C-J, Yeh C-Y, Yang W-CV, et al. Heterogeneity of Alzheimer's disease identified by neuropsychological test profiling. Plos One. 2023;18: e0292527.

25. Bombina P, Tally D, Abrams ZB, Coombes KR. SillyPutty: Improved clustering by optimizing the silhouette width. bioRxiv. 2023.

26. Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology. 1993.

27. Kaplan E, Goodglass H, Weintraub S. Boston naming test. Clin Neuropsychol. 2001.

28. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12: 189–198.

29. Ball LJ, Bisher GB, Birge SJ. A Simple Test of Central Processing Speed: An Extension of the Short Blessed Test. J Am Geriatr Soc. 1999;47: 1359–1363. doi:10.1111/j.1532-5415.1999.tb07440.x

30. Watkins MJ, Neath I, Sechler ES. Recency Effect in Recall of a Word List When an Immediate Memory Task Is Performed after Each Word Presentation. Am J Psychol. 1989;102: 265–270. doi:10.2307/1422957

31. Shao Z, Janse E, Visser K, Meyer AS. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. Front Psychol. 2014;5. doi:10.3389/fpsyg.2014.00772

32. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9.

33. Zhou Y-H, Saghapour E. ImputEHR: a visualization tool of imputation for the prediction of biomedical data. Front Genet. 2021;12: 691274.

34. Batool F, Hennig C. Clustering with the Average Silhouette Width. Comput Stat Data Anal. 2021;158: 107190. doi:10.1016/j.csda.2021.107190

35. SaeedNasir, NamHaewoon, Ul HI, Bhatti MS. A Survey on Multidimensional Scaling. ACM Comput Surv CSUR. 2018 [cited 3 Jun 2024]. doi:10.1145/3178155

Acknowledgements

The preparation of this report was supported by the Centene Corporation contract (P19-00559) for the Washington University-Centene ARCH Personalized Medicine Initiative.

Competing interests

No authors have any conflicting interests to report.

Author Contributions

All authors contributed to the design of the methodology and the experiments. SK implemented the data pre-processing, modelling, and data analysis and manuscript writing. ZA provided guidance on data analysis, model building, and validation. ZA, PP, SS, and NG provided conceptualization, clinical interpretation, and manuscript writing. All authors participated in manuscript revision and editing.

Data availability and consent statement

Anonymized data not published within this article will be made available by request from any qualified investigator. Approval for this study was obtained from the Washington University Institutional Review Board, which granted a waiver of HIPAA Authorization for the use of Protected Health Information (PHI).

Figures and Tables

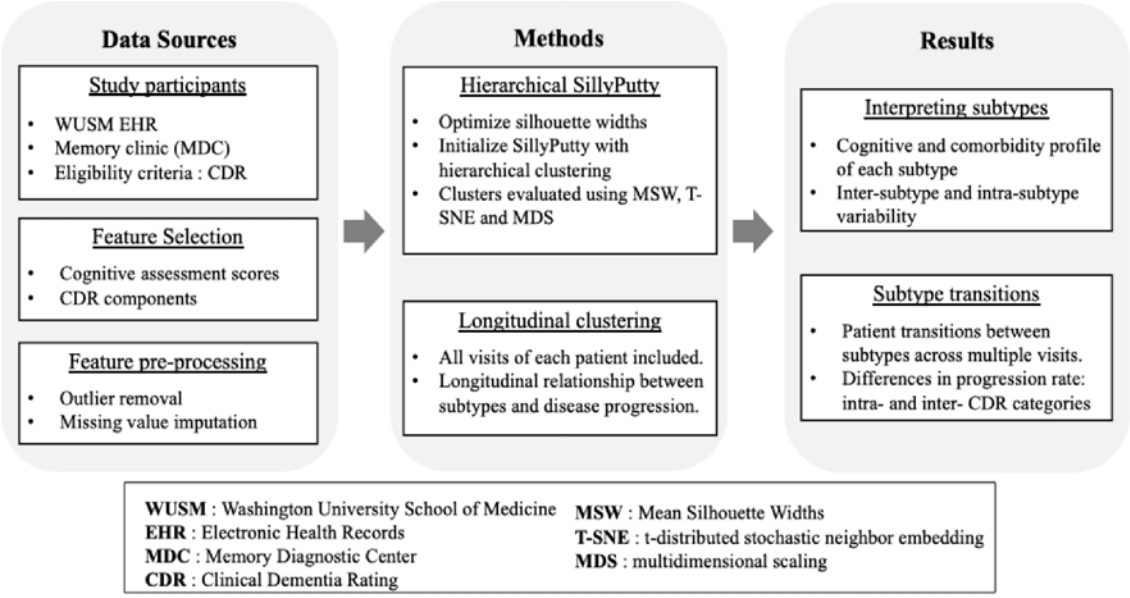


Figure 1: Study workflow. Outpatient office visits from Washington University School of Medicine (EHR), with recorded CDR were included. Feature selection performed to identify the most optimal set of features for clustering. SillyPutty algorithm[25], initialized with hierarchical clustering was used to identify clusters from longitudinal visits of patients. Identified clusters were analyzed for variability in cognitive and comorbidity characteristics. Patient transitions between subtypes across multiple visits were examined for variability in rate of disease progression.

Table 1: Dementia cohort demographics extracted from the EHR

Variable	Total
Number of patients	1845
Age at first encounter, median (IQR), years	73 [64-81]
Sex, N (%) - Female	 1038 (57%)
Race, N (%) - White - Black or African American - Asian - *Other	 1422 (77%) 181 (9.8%) 20 (1.2%) 222 (12%)
Ethnicity, N (%) - Non-Hispanic or Latino - Hispanic or Latino - Unknown	 1643 (88.5%) 84 (5.1%) 118 (6.4%)
Most Common Diagnoses (n, %) -Memory Loss -Alzheimer Disease -Parkinsonism -Major Depressive Disorder -Obstructive Sleep Apnea	 987 (54 %) 1061 (57.5 %) 66 (3.6 %) 455 (24.7%) 366 (19.8%)

* Other includes Native Hawaiian or Other Pacific Islander, Other, Unknown, Declined, or unreported.

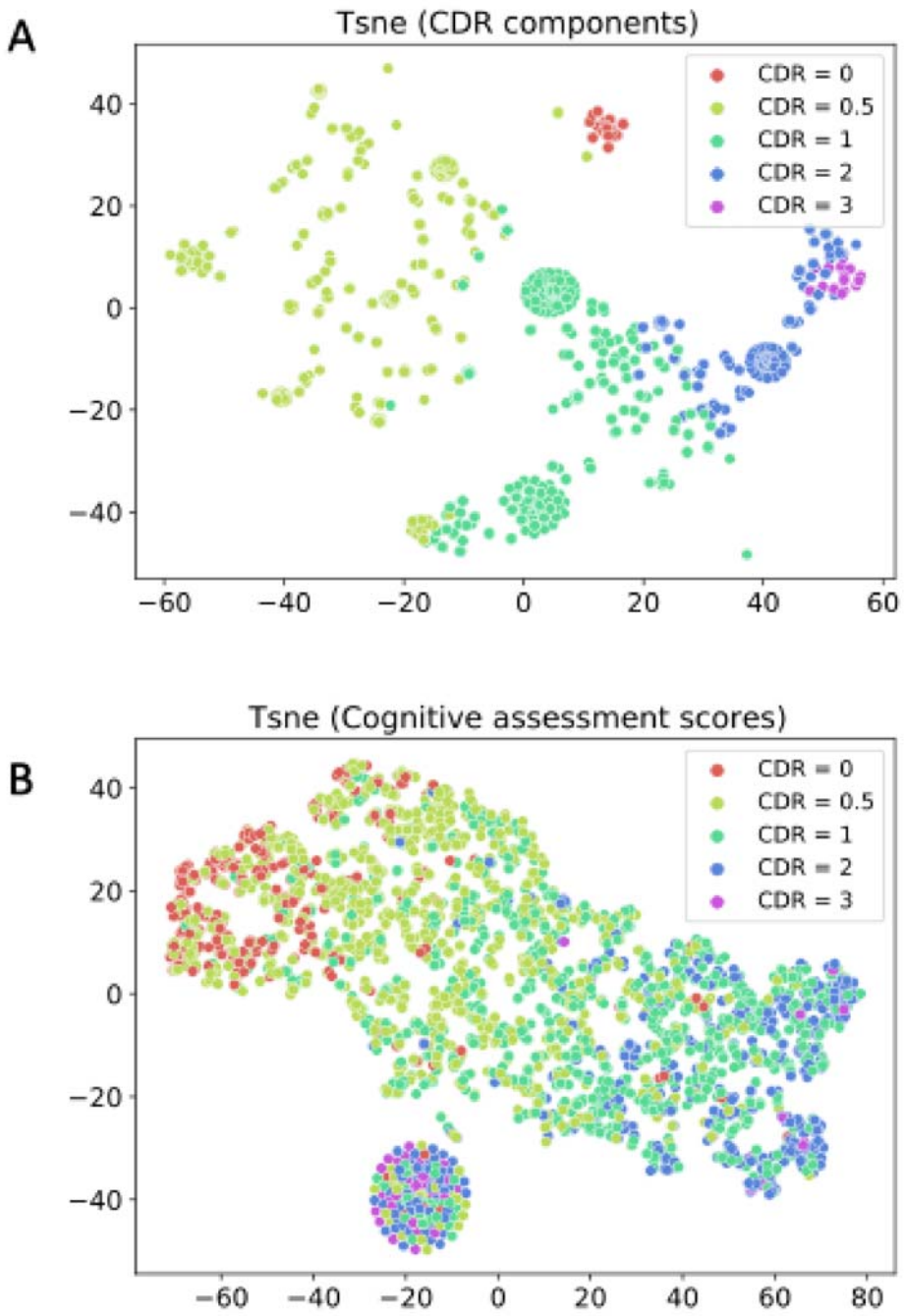


Figure 2: 2D T-SNE representations of the data for each of the 2 feature categories: cognitive scores only as features (**2A**) and CDR components only as features (**2B**). Each point in the scatter plot represents a visit with the colour indicating the CDR score. The x-axis and y-axis represent the 2 dimensions of the 2D T-SNE vector for visualization purposes.

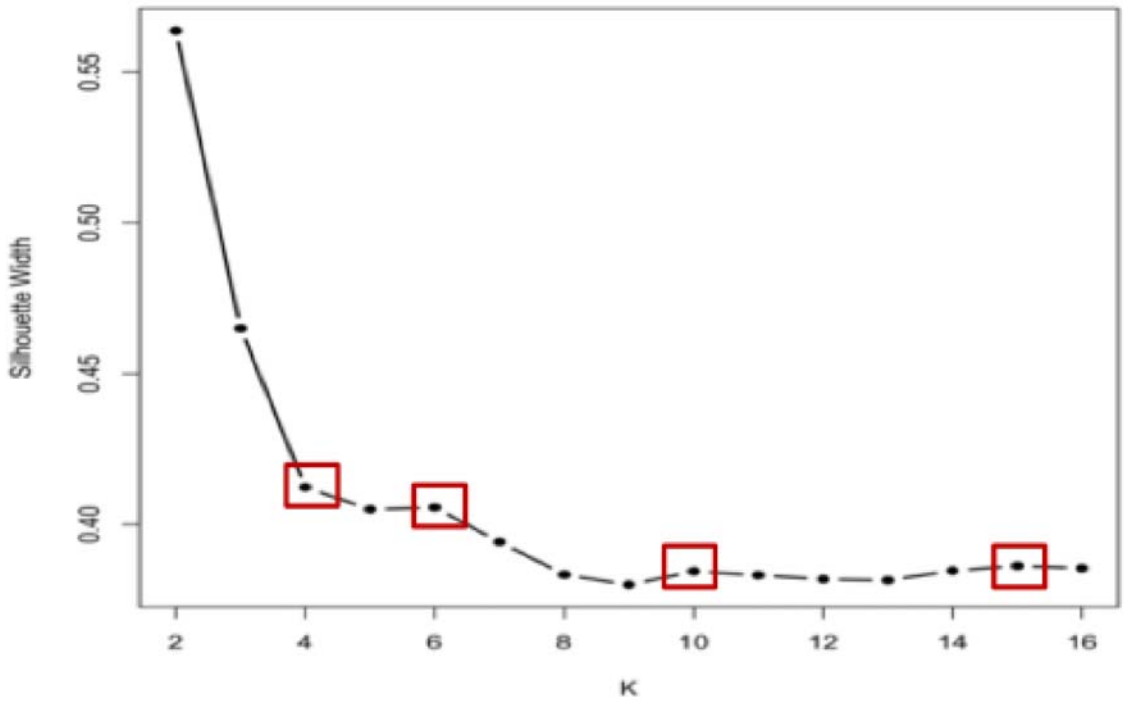


Figure 3: Optimal number of clusters for Hierarchical SillyPutty were determined by comparing the best mean silhouette width (MSW), for a range of number of clusters (K) from 2 to 16. K = 10 was selected as the optimal number of clusters for our analyses.

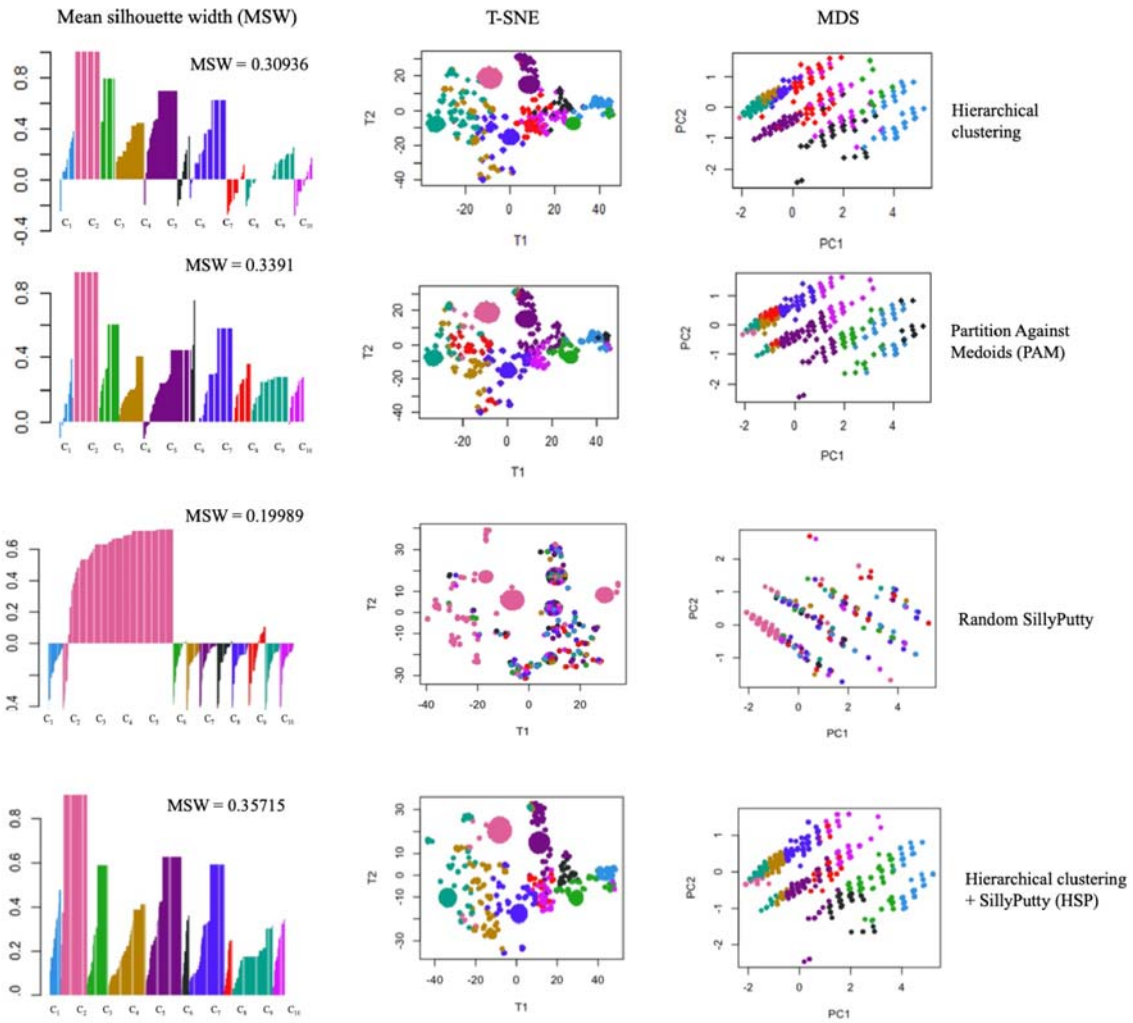


Figure 4: Clustering statistics of different clustering algorithms: hierarchical clustering, PAM, random SillyPutty and our method hierarchical SillyPutty (from top to bottom). For each clustering method, the plots from left to right represent mean silhouette width (MSW), T-sne distribution and multidimensional scaling (MDS) respectively.

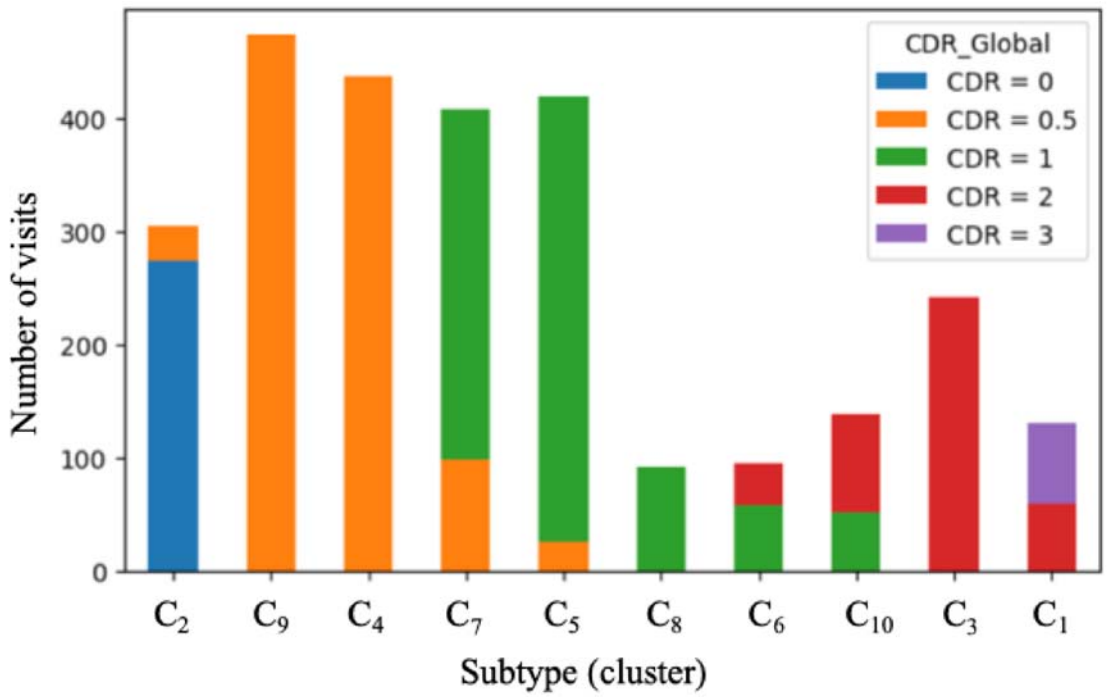


Figure 5: Stacked bar plot showing the CDR composition of each dementia subtypes ordered by increasing CDR score (dementia severity increases moving from left to right). The x-axis shows the 10 dementia subtypes, and the y-axis represents the number of visits within each dementia subtypes. Some dementia subtypes included a unique CDR (e.g. C_9, C_4), while other dementia subtypes included two CDR scores (e.g. C_7, C_5).

Table 2: Cluster demographics based on baseline visits of each visits of each patient

Variable	**C_2**	**C_9**	**C_4**	**C_7**	**C_5**	**C_8**	**C_6**	**C_{10}**	**C_3**	**C_1**
Number of patients	194	388	281	293	301	47	51	73	158	59
Number of visits	302	495	436	409	426	87	91	130	225	118
Age, median (IQR)	69 (13)	73 (12)	76 (11)	76 (13)	75(12)	77 (11)	77 (15)	77 (11)	76 (12)	72 (14)
Sex, (%) - Female	54.3%	64.2%	52.6%	62.7%	48.4%	51.2%	49.8%	65%	63.6%	72.5%
Race, (%) - White - Black - Asian	 90.6% 8.5% 0.82%	 91.2% 7.7% 1.2%	 89.3% 9.4% 1.3%	 89.4% 9.7% 0.9%	 91.8% 8.5% 0.82%	 83.3% 15.7% 0.93%	 94% 6% 0%	 83.7% 15.4% 0.85%	 88.2% 10.8% 0.82%	 90.6% 8.5% 0.98%
Brain disorders, (%) - Memory Loss - Alzheimer Disease - Parkinsonism - Depression - Sleep disorders	 86.5% 6.4% 0.46% 3.2% 3.4%	 64.2% 19.2% 1.3% 1.4% 2.9%	 57.6% 24.5% 1.36% 0.81% 1.9%	 52.3% 44.6% 1.2% 0.62% 1.2%	 57.5% 40% 1.6% 1.25% 0%	 75% 25% 0% 0% 0%	 58.8% 38.2% 30.6% 4.1% 1.1%	 38.2% 54.2% 0.8% 0% 1%	 29.7% 67.2% 2.1% 0% 1.4%	 31.5% 63.5% 1.4% 0% 4.5%
Cognitive assessment scores, median (IQR) - Boston Naming Test - MMSE - Short Blessed - Word List Memory - Verbal Fluency	 15 (1) 28 (15) 2 (4) 18 (7) 16 (7)	 14 (2) 25 (14) 6 (9) 13 (5) 12 (6)	 14 (4) 22 (14) 12 (11) 12 (6) 11 (6)	 13 (3) 22 (13) 12 (10) 12 (4) 11 (5)	 13 (4) 19 (13) 18 (13) 11 (6) 9 (6)	 14 (4) 20 (13) 16 (14) 12 (5) 9 (6)	 13 (4) 18 (12) 17 (11) 9 (5.7) 8 (6)	 11 (6) 15 (11) 21 (8) 10 (7) 8 (7)	 12 (5) 16 (14) 21 (16) 13 (7) 9 (7)	 14 (4.5) 23 (13) 10 (14) 13 (5) 12 (4)

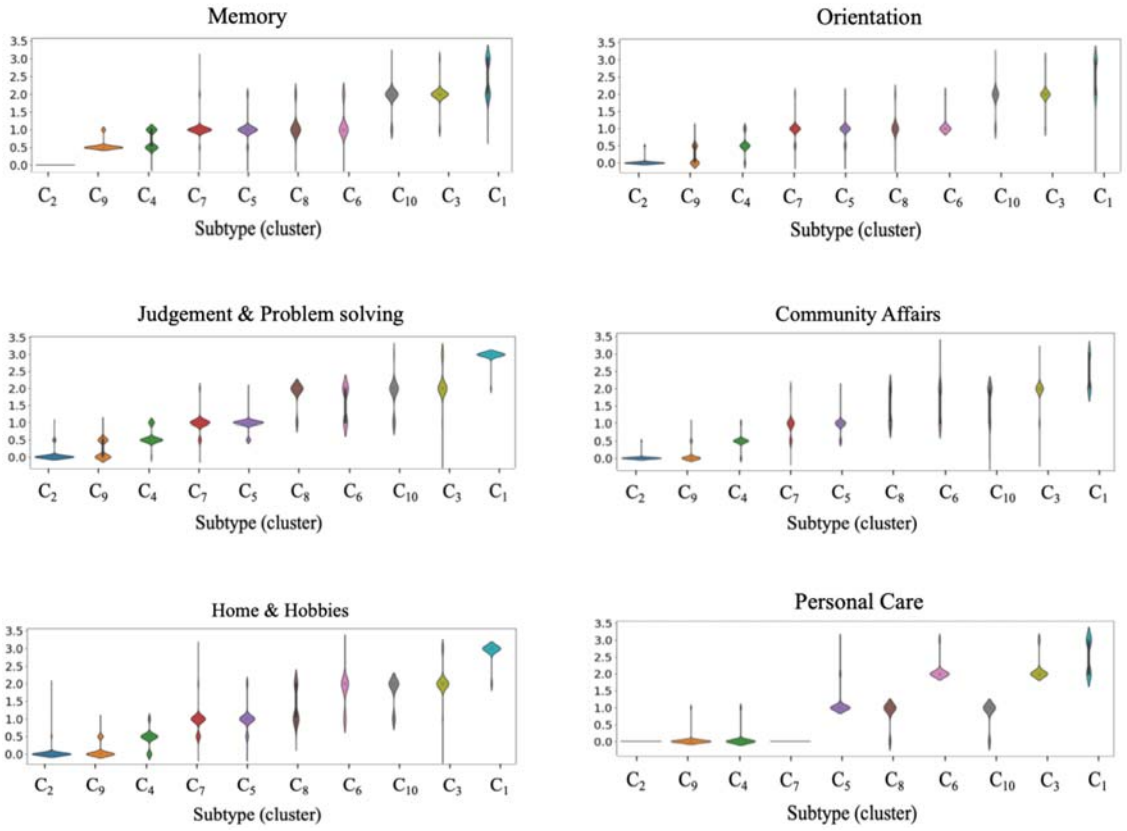


Figure 6: Violin plots showing how each of the how each of the 6 components of CDR score vary across the 16 dementia subtypes. The x-axis represents the individual dementia subtypes. The y-axis shows the six CDR component scores varying from 0-3, with 0 indicating normal cognition and 3 indicating severe impairment.

Table 3: Cognitive profile summary (CDR components) of each subtype. Subtypes have been grouped based on similarity in their cognitive characteristics.

Subtype	CDR	Cognitive profile summary
C_2	0, 0.5	Relatively healthy population. Minimal cognitive and functional impairment.
C_9	0.5	Mild to moderate memory loss, unimpaired functional capacity in other cognitive and functional domains.
C_4	0.5	Mild to moderate memory loss. Slight disorientation, and slight impairment in judgement, community, and home activities
C_7 C_5	0.5, 1	Moderate impairment in all six domains. Patients in C_7 were fully capable of self-care activities.
C_8 C_6	1 1, 2	Moderate memory and orientation problems. Severe impairment in judgement, community, and home activities.
C_{10} C_3 C_1	1, 2 2 2, 3	Severe impairment in all 6 cognitive and functional domains with maximum impairment level in C_1. C_{10} showed moderate difficulty in personal care with need for caregiving support.

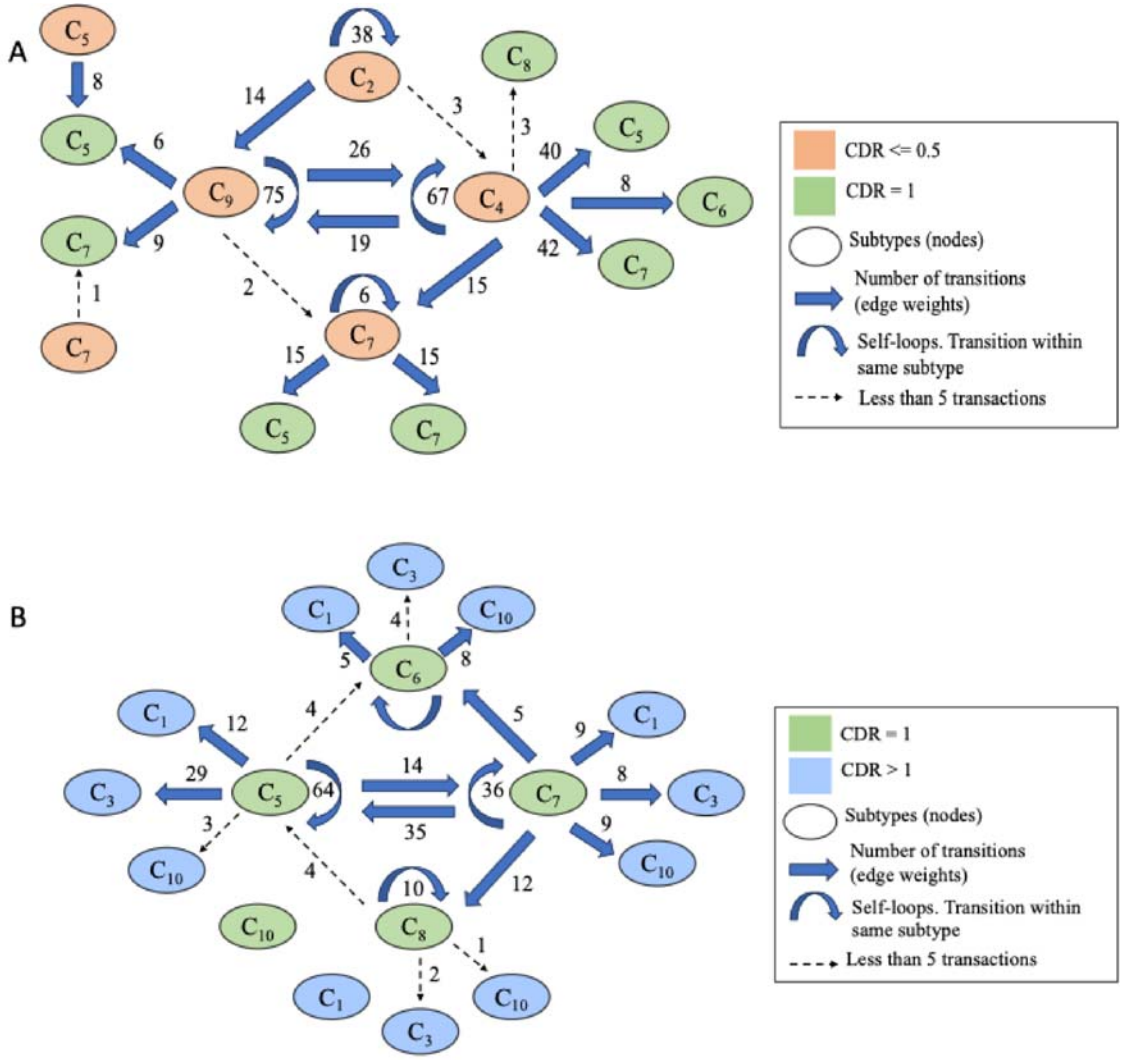
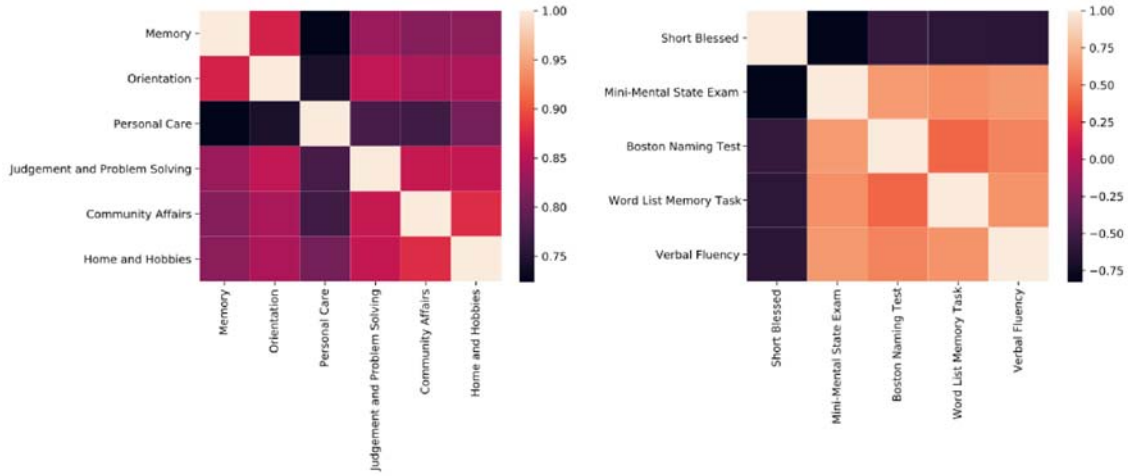
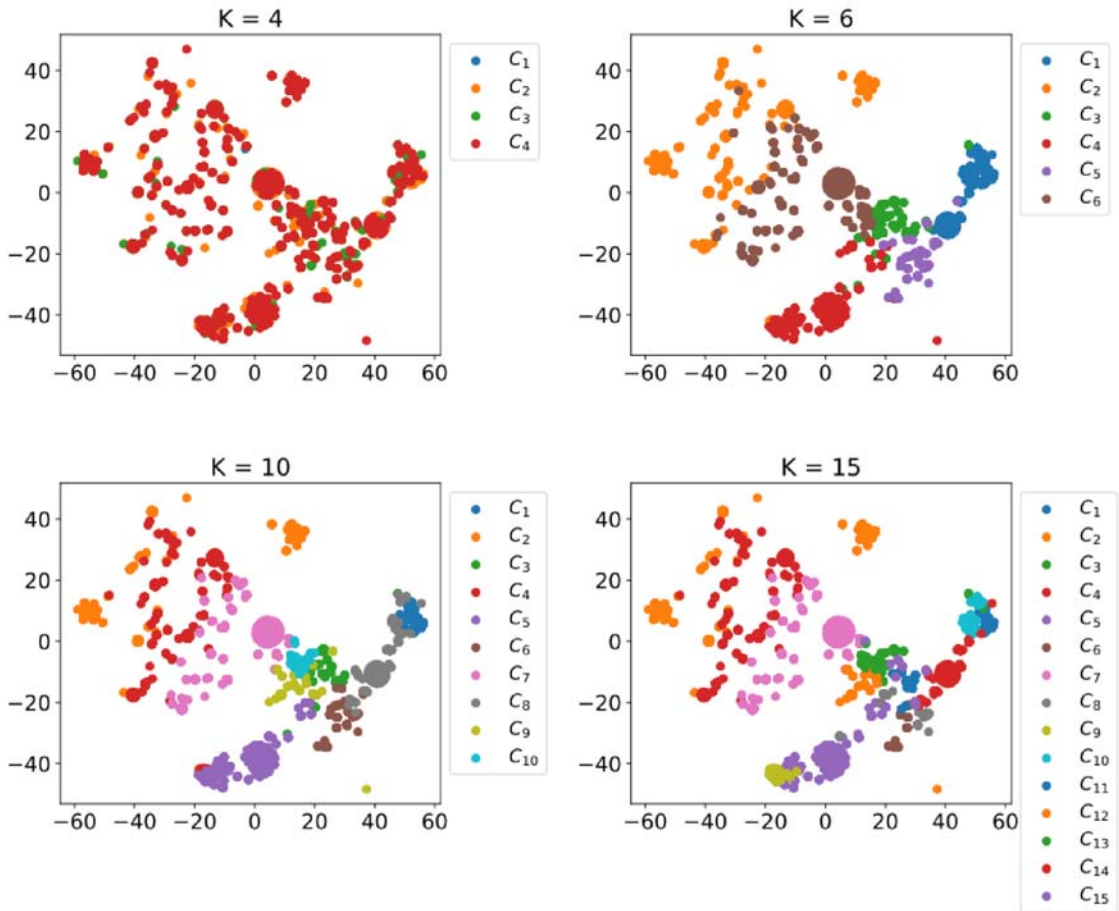


Figure 7: Graph showing transitions between subtypes which are related to CDR progression from CDR <= 0.5 to CDR = 1 (**A**) and CDR = 1 to CDR > 1 (**B**). The nodes represent the subtypes, and the edge weights represent the number of transitions from source to target. Transitions with edge weights >= 5 are shown with dotted arrows. Transitions between nodes of same color including self-transitions indicate transitions remaining in the same CDR stage.

Supplementary Figures



eFigure 1: Correlation heatmap of features for (**A**) CDR component scores and (**B**) cognitive assessment scores.



eFigure 2: T-SNE distribution of the clustering results for different values of K = 4, 6, 10 and 15 respectively. Each point in the scatter plot represents a single visit. The x-axis and y-axis represent the 2 dimensions of the 2D T-SNE vector for visualization purposes.

eTable 1: Number of transitions between every pair of dementia subtype groups. The subtype groups in 1st column and 1st row represent source and target respectively. The last column represents the total number of transitions from each source subtype.

Start/End	2	9	4	7	5	8	6	10	3	1	Total
2	38	14	3	0	0	0	0	0	1	0	56
9	15	75	24	13	7	1	0	1	0	0	136
4	2	19	67	56	44	3	5	0	3	0	199
7	0	1	7	54	50	15	5	9	8	9	158
5	0	2	5	16	75	5	4	8	29	12	156
8	0	0	0	8	4	10	2	2	1	0	27
6	0	0	0	0	1	3	13	8	4	5	34
10	0	0	0	1	2	1	6	12	8	7	37
3	0	0	0	0	0	0	1	9	37	19	66
1	0	0	0	0	0	0	0	1	3	19	23

Data Sources

Study participants

- WUSM EHR
- Memory clinic (MDC)
- Eligibility criteria : CDR

Feature Selection

- Cognitive assessment scores
- CDR components

Feature pre-processing

- Outlier removal
- Missing value imputation

Methods

Hierarchical SillyPutty

- Optimize silhouette widths
- Initialize SillyPutty with hierarchical clustering
- Clusters evaluated using MSW, T-SNE and MDS

Longitudinal clustering

- All visits of each patient included.
- Longitudinal relationship between subtypes and disease progression.

Results

Interpreting subtypes

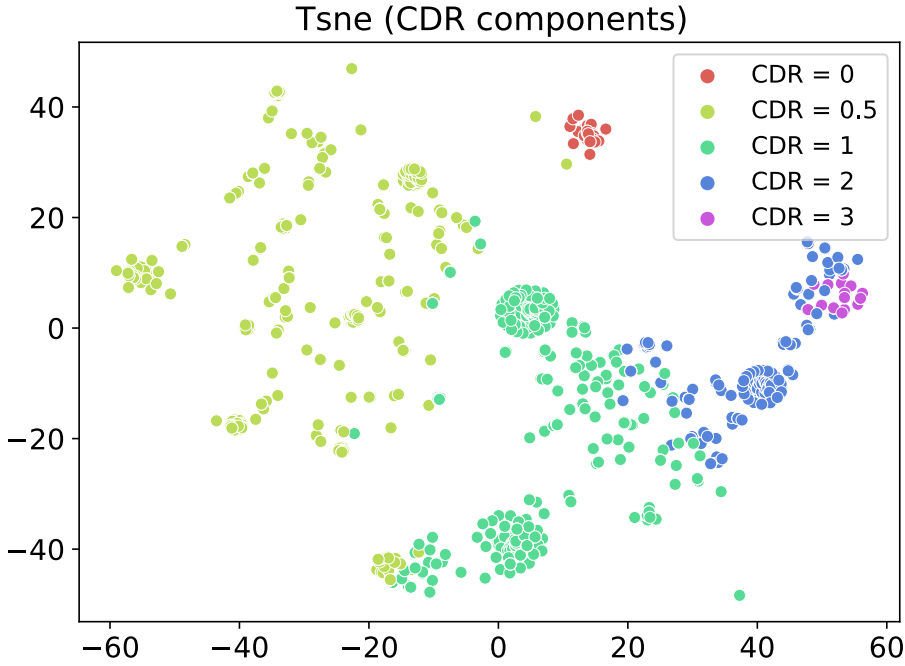
- Cognitive and comorbidity profile of each subtype
- Inter-subtype and intra-subtype variability

Subtype transitions

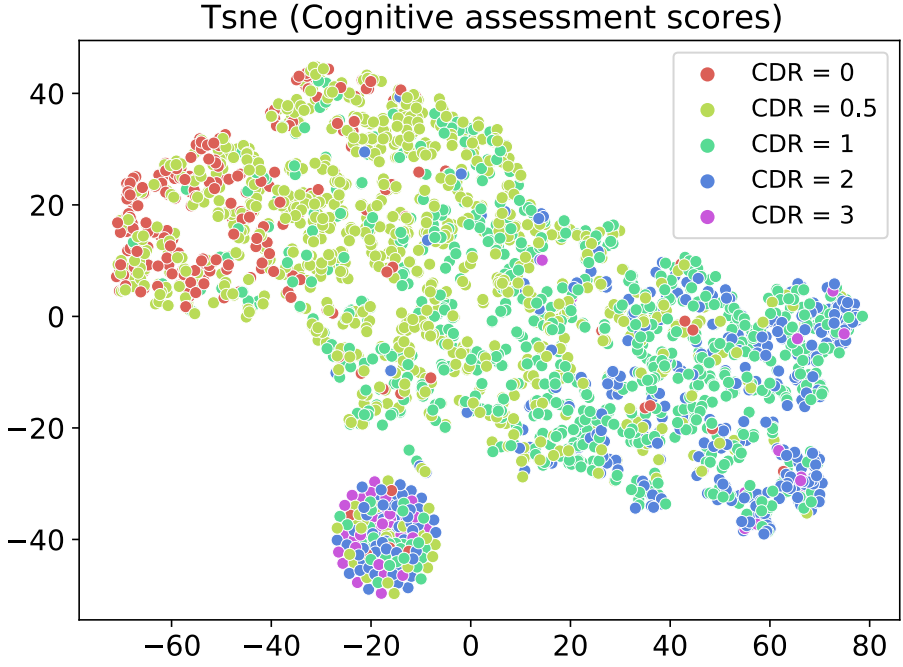
- Patient transitions between subtypes across multiple visits.
- Differences in progression rate: intra- and inter- CDR categories

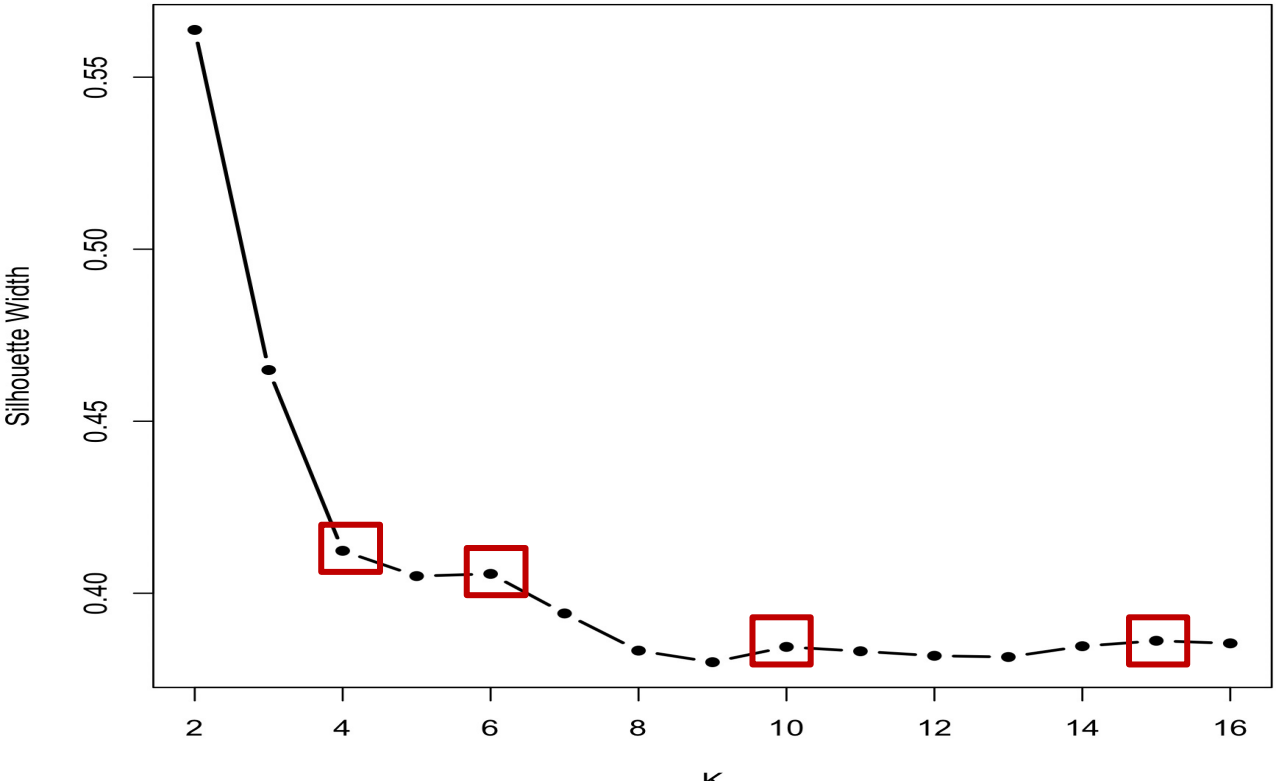
WUSM : Washington University School of Medicine
EHR : Electronic Health Records
MDC : Memory Diagnostic Center
CDR : Clinical Dementia Rating
MSW : Mean Silhouette Widths
T-SNE : t-distributed stochastic neighbor embedding
MDS : multidimensional scaling

A


Tsne (CDR components)
CDR = 0
CDR = 0.5
CDR = 1
CDR = 2
CDR = 3
40
20
0
−20
−40
−60
−40
−20
0
20
40
60

B

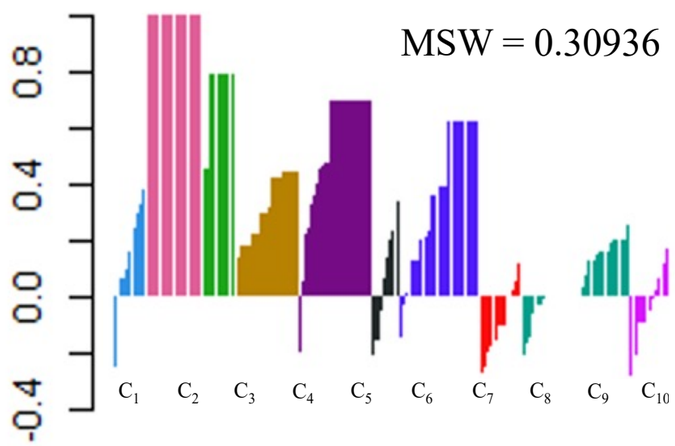

Tsne (Cognitive assessment scores)
CDR = 0
CDR = 0.5
CDR = 1
CDR = 2
CDR = 3
40
20
0
−20
−40
−60
−40
−20
0
20
40
60
80

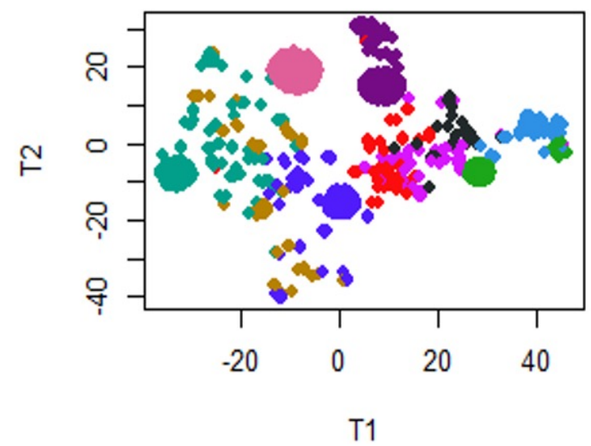

Silhouette Width
0.40
0.45
0.50
0.55
2
4
6
8
10
12
14
16
K

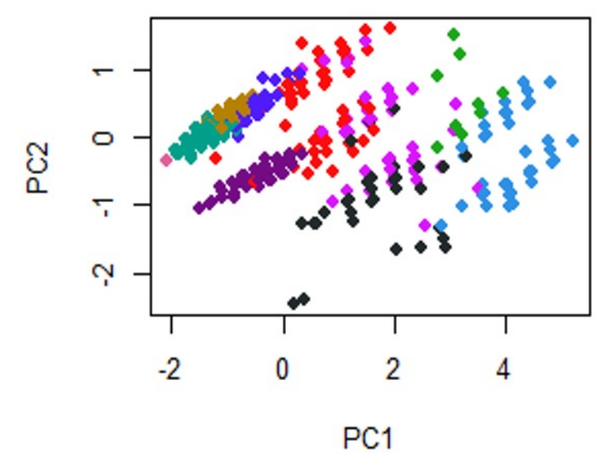
Mean silhouette width (MSW)

T-SNE

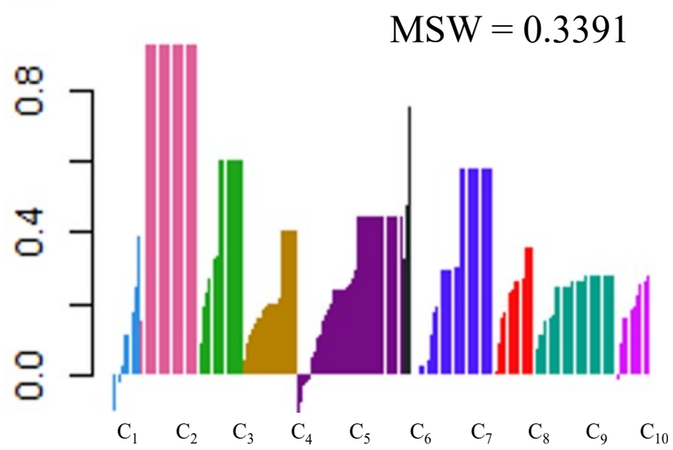
MDS

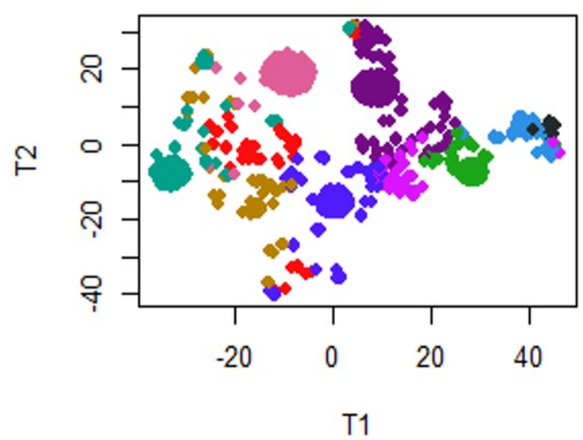


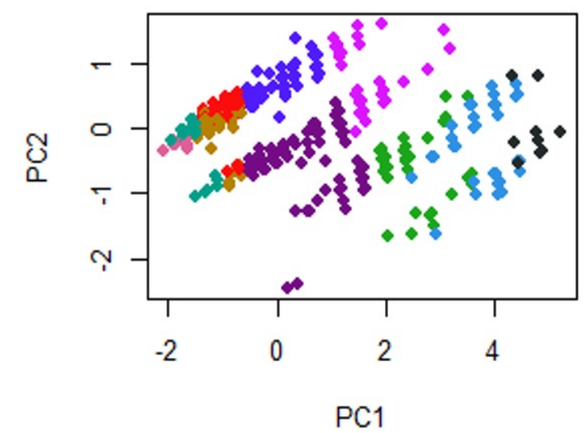




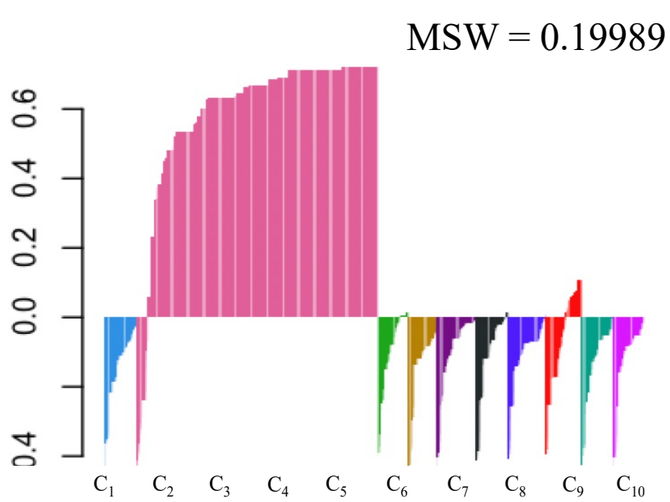
Hierarchical clustering

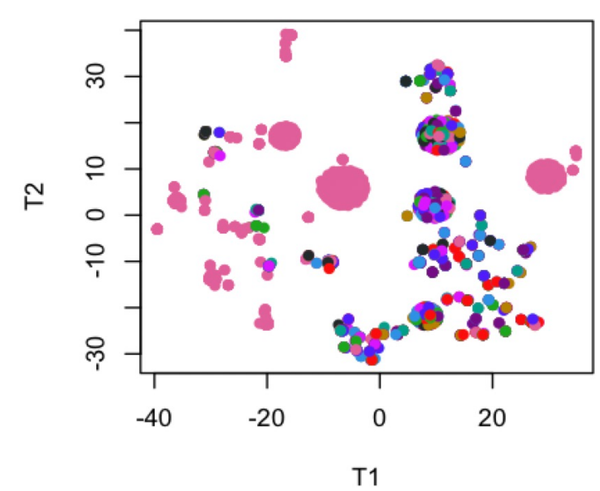


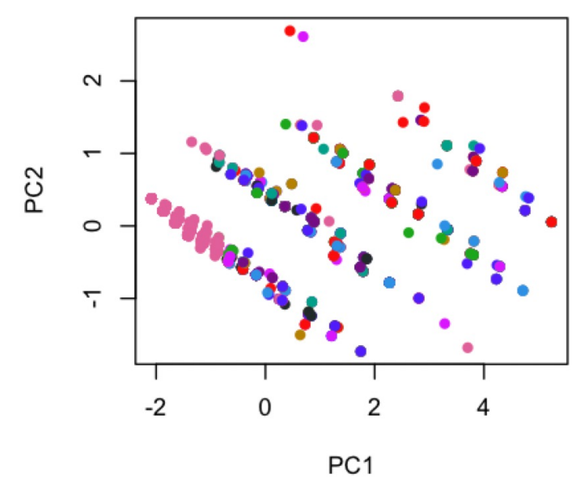




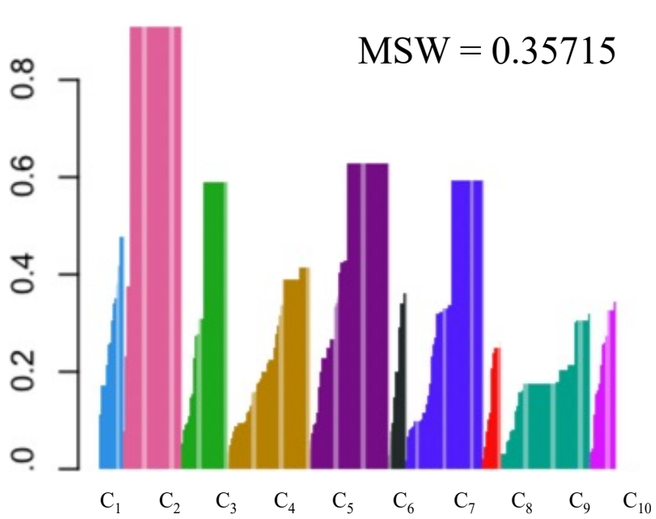
Partition Against Medoids (PAM)

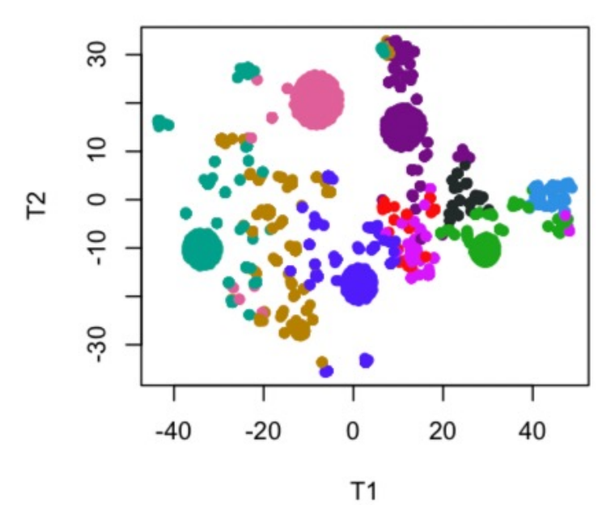


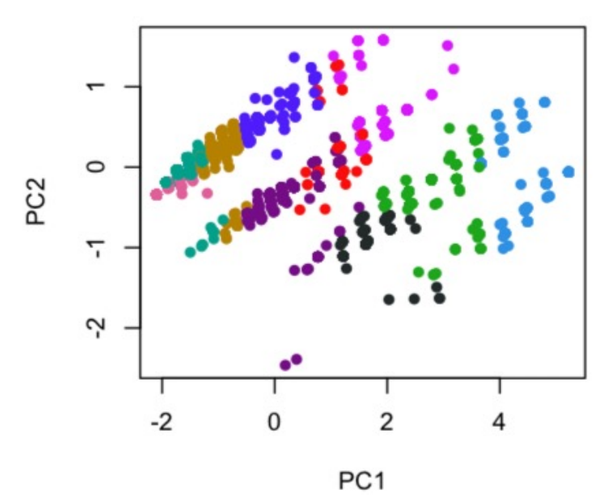




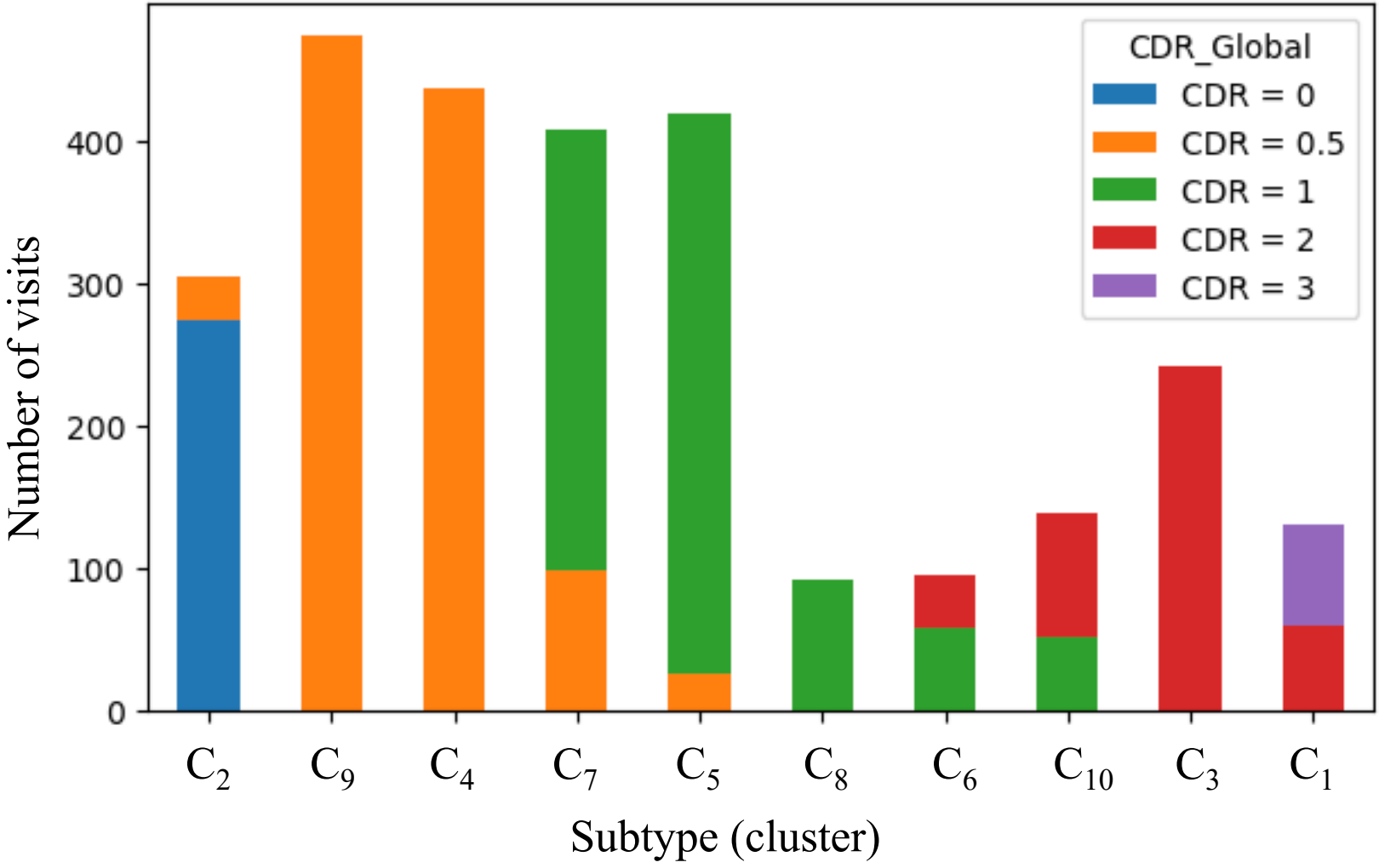
Random SillyPutty



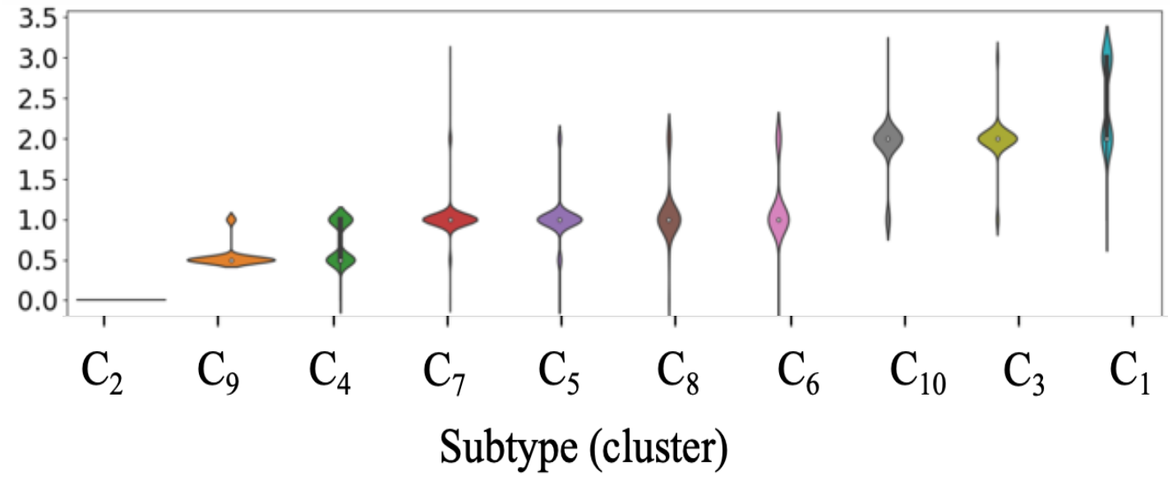




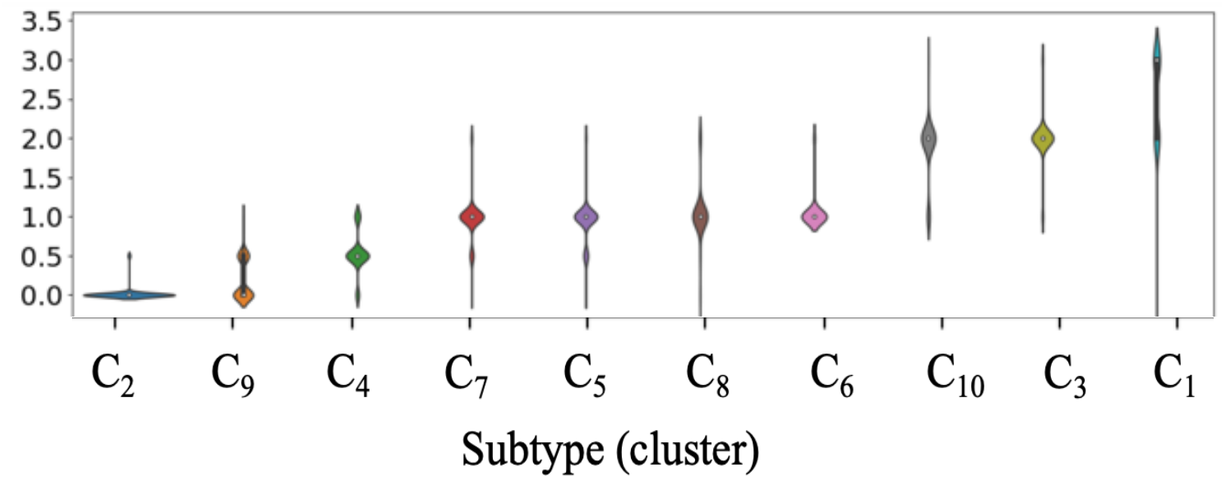
Hierarchical clustering + SillyPutty (HSP)


CDR_Global
CDR = 0
CDR = 0.5
CDR = 1
CDR = 2
CDR = 3
Number of visits
0
100
200
300
400
C2
C9
C4
C7
C5
C8
C6
C10
C3
C1
Subtype (cluster)

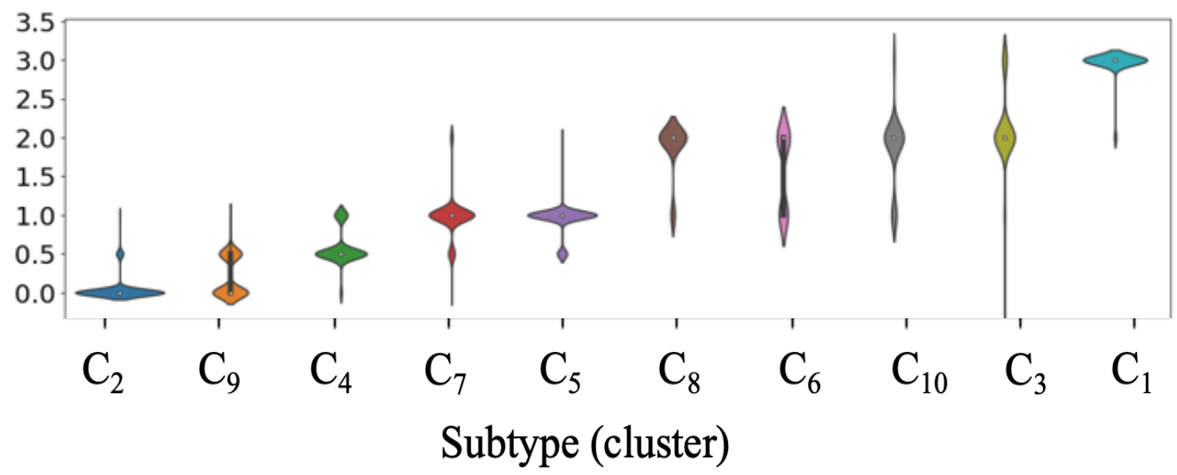
Memory



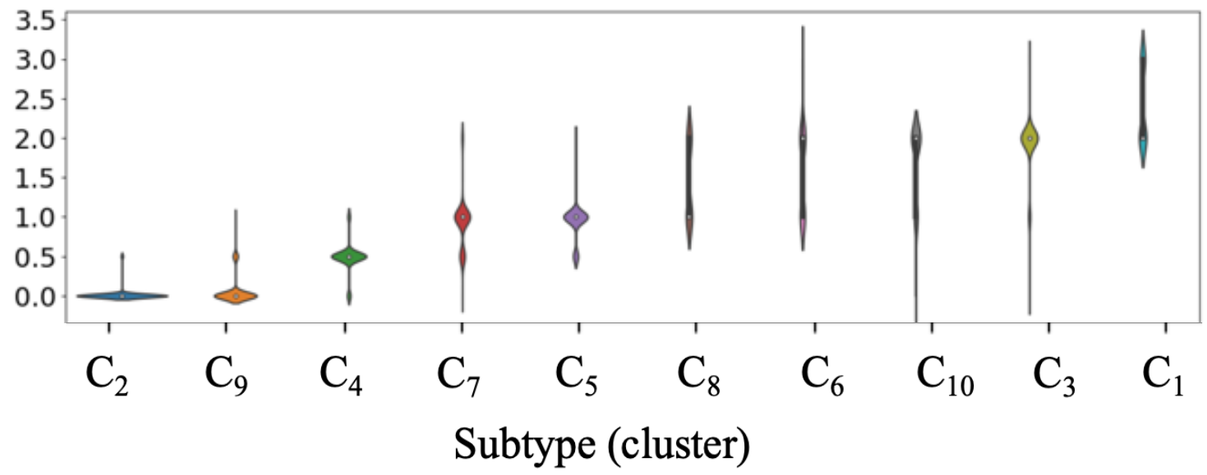
Orientation



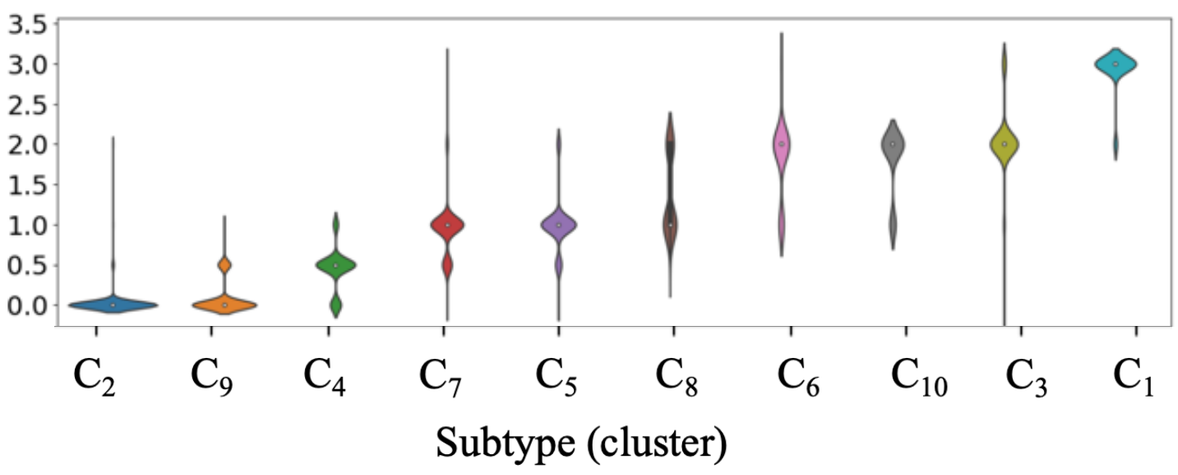
Judgement & Problem solving



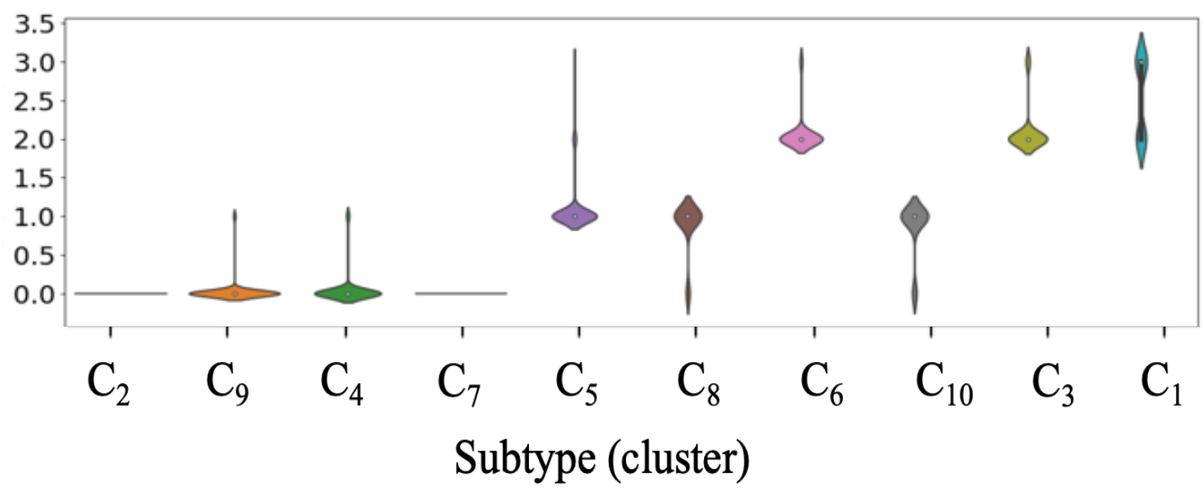
Community Affairs

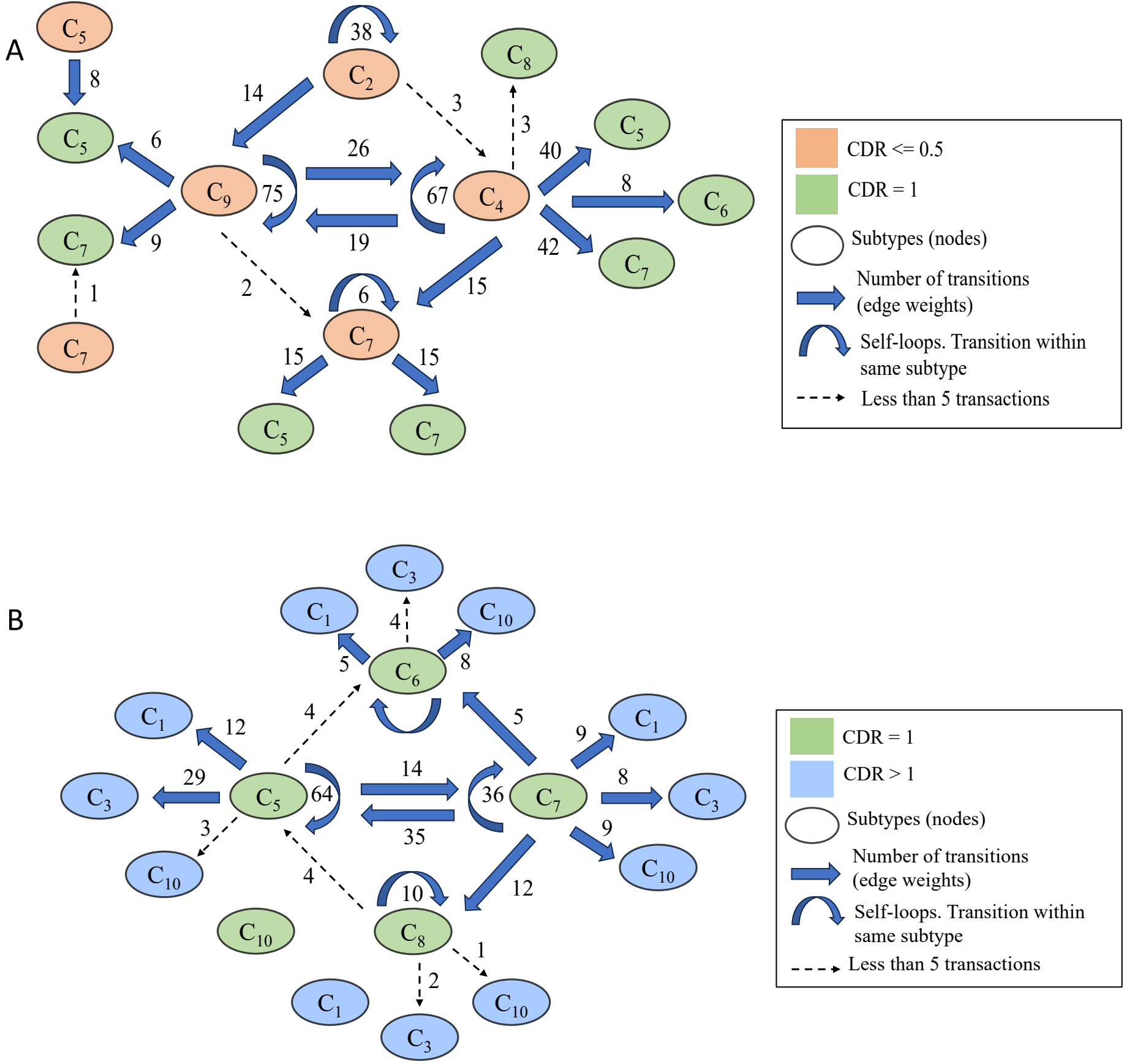


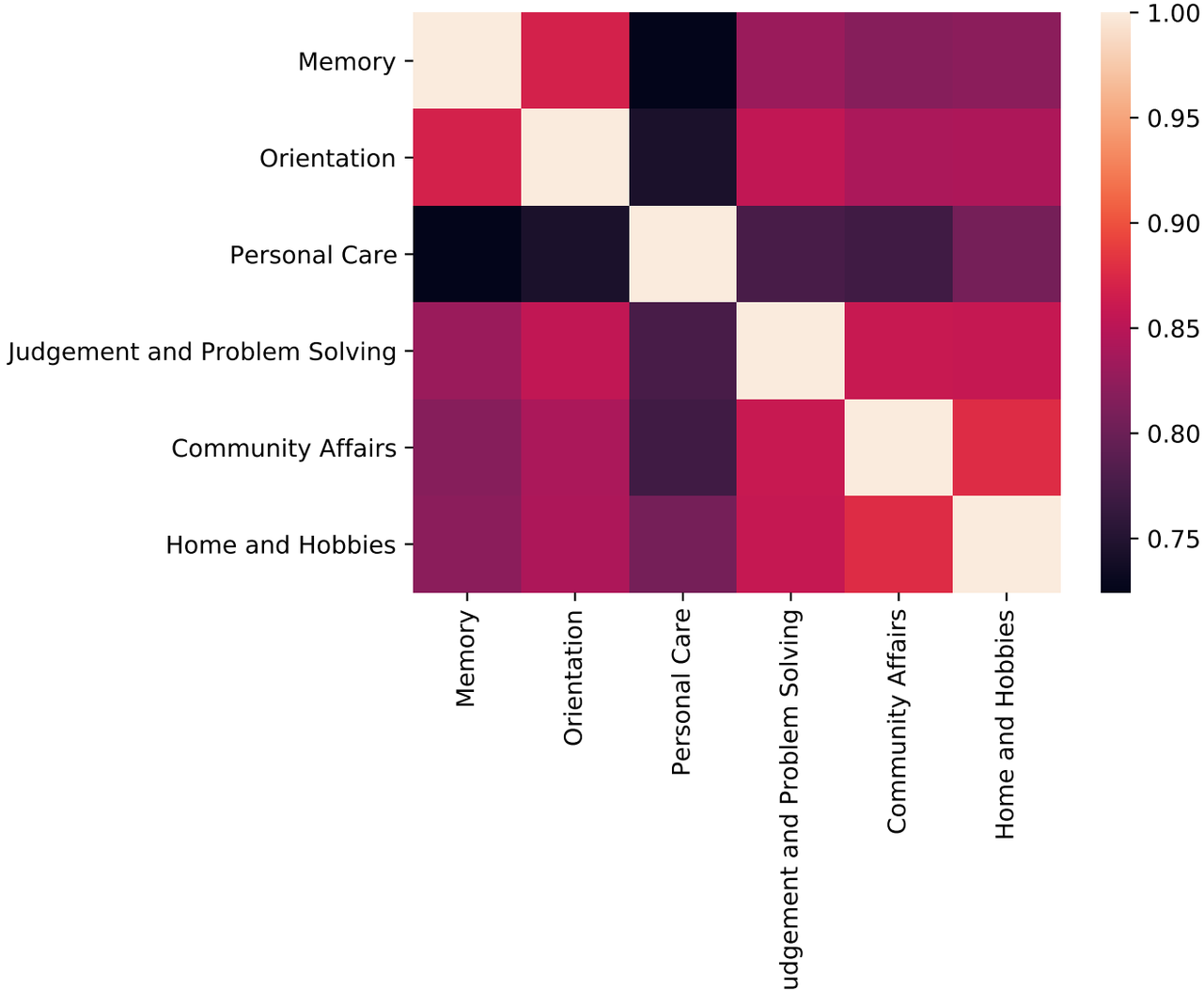
Home & Hobbies

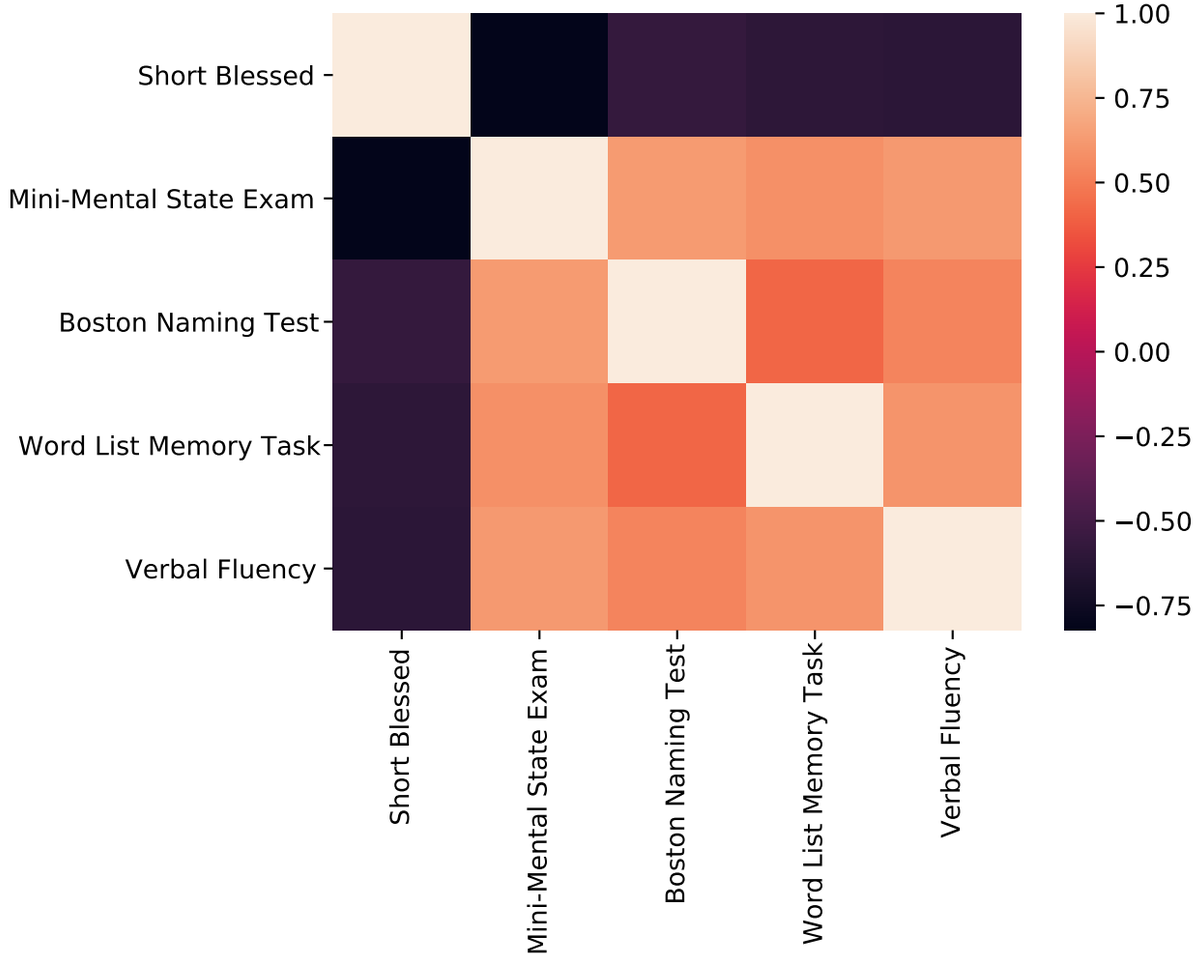


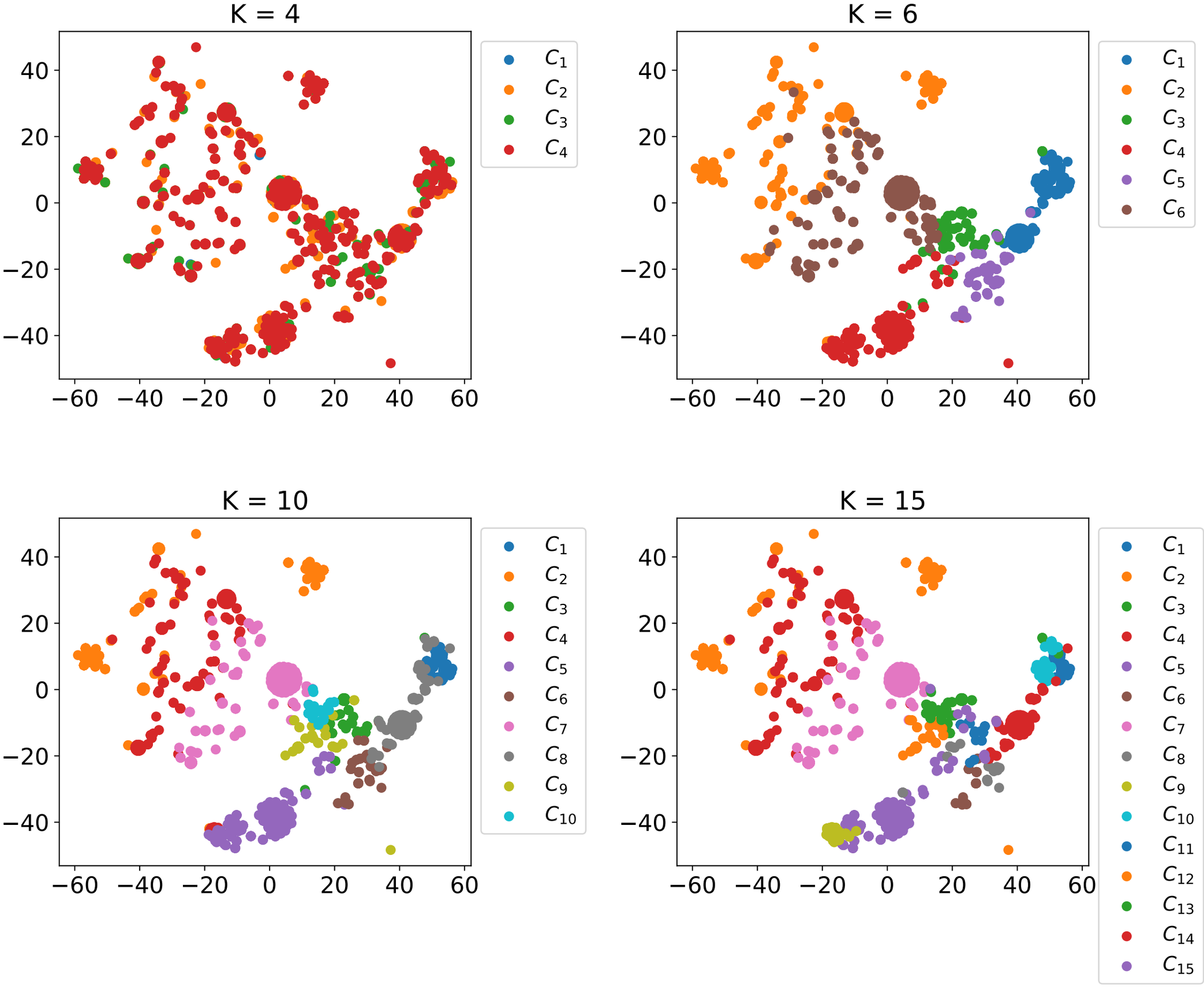
Personal Care




A
C5
8
C5
6
C9
75
26
19
C2
38
14
3
C8
3
67
C4
40
C5
8
C6
42
C7
C7
9
1
C7
2
6
C7
15
15
15
C5
C7
CDR <= 0.5
CDR = 1
Subtypes (nodes)
Number of transitions (edge weights)
Self-loops. Transition within same subtype
Less than 5 transactions
B
C3
C1
C10
4
5
C6
8
C1
12
4
5
9
C1
29
C3
C5
64
14
36
C7
8
C3
3
35
9
C10
C10
4
10
12
C10
C8
1
2
C1
C10
C3
CDR = 1
CDR > 1
Subtypes (nodes)
Number of transitions (edge weights)
Self-loops. Transition within same subtype
Less than 5 transactions


Memory
Orientation
Personal Care
Judgement and Problem Solving
Community Affairs
Home and Hobbies
1.00
0.95
0.90
0.85
0.80
0.75


Short Blessed
Mini-Mental State Exam
Boston Naming Test
Word List Memory Task
Verbal Fluency
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


K = 4
K = 6
K = 10
K = 15
C_1
C_2
C_3
C_4
C_5
C_6
C_7
C_8
C_9
C_{10}
C_{11}
C_{12}
C_{13}
C_{14}
C_{15}
40
20
0
−20
−40
−60
−40
−20
0
20
40
60