# Supplementary Materials

## Clustering cognitive phenotypes in affective and non-affective psychosis

Katharina M Bracher[1], Afra Wohlschläger[2], Kathrin Koch[2], Franziska Knolle[2]*

[1]) Division of Neurobiology, Faculty of Biology, LMU Munich, Martinsried 82152, Germany; Graduate School of Systemic Neurosciences, LMU Munich, Martinsried 82152, Germany

[2]) Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Technical University of Munich, Munich, Germany

* corresponding author: franziska.knolle@tum.de

Conflict of Interest: None of the authors declares a conflict of interest.

Funding: FK received funding from the European Union's Horizon 2020 [Grant number 754462].

# Contents

# 1 Data

## 1.1 Features Selection

From the demographic, behavioural and psychological data, we selected 160 features as potentially relevant to our analysis (Suppl. Tab. 2, 3). Features containing information of primary diagnosis for affective and non-affective psychosis, such as the Positive and Negative Symptom Score (panss01) and the Clinical Assessment Interview for Negative Symptoms (cains01) were not included in the analysis. Furthermore, we dropped 105 features (Suppl. Tab. 2) that had too few entries (more than 80% missing entries), which resulted in the selection of 54 final features (Suppl. Tab. 3 for a complete list of all features and tests). Many of the 54 selected features contained missing data, hence, the goal was to obtain a data-set with as many subjects as possible, allowing for a similar distribution of controls and patients, while keeping the number of missing values small. The optimal trade-off was achieved when excluding subjects that were missing more than four entries of the selected features. Using this trade-off, we were able to maintain a high number of subjects - 217 (55 healthy controls, 51 affective psychosis group, 111 non-affective psychosis group) of the original 251 (68 healthy controls, 57 affective psychosis group, 126 non-affective psychosis group). The distribution of the remaining subjects reflected the distribution of the original data.

For those individuals who had a sufficient number of entries, missing data were imputed using the mean for continuous and mode for categorical features (Suppl. Tab. 1). Since the ratio of patients and controls was not balanced, mean and mode were calculated separately for patients and controls. Both patient groups were combined in order to minimise the bias of classical group membership based on diagnosis regarding cognitive impairments. The maximal number of missing entries was 25 of 217 subjects for "age-adjusted fluid cognition composition score" (Suppl. Tab. 1).

The released brain imaging data contain structural magnetic resonance imaging (MRI), resting state functional MRI, and diffusion MRI data. Here, we selected the T1-weighted structural image of all 183 individuals. Structural images were recorded at a 3T SIEMENS MAGNETOM Prisma scanner using a MPRAGE sequence (TR=2400ms, TE=2.22ms, FoV read=256mm, FoV phase=93.8%, flip angle=8 deg, slices per slap=208, slice thickness=0.8mm).

Suppl. Table 1: Features with number of missing individuals. Missing values for continuous features were replaced by the mean according to the group of patients or controls. For categorical data, the mode was used to replace the missing value. *W* stands for 'white', for Socio-Economic Status, 2 for a SES score of 20 to 29 and in the Mother/Father Educational Scale a 4 for High School Graduation or GED and a 6 for Completed 7th through the 9th grades.

| Feature | Missing | Controls mean(std)/mode | Patients mean(std)/mode |
|---|---|---|---|
| **Auditory continuous performance test** | | | |
| auditory_t2 | 12 | -0.48 (0.87) | 0.15 (0.99) |
| auditory_t4 | 2 | 0.39 (0.59) | -0.13 (1.08) |
| auditory_t5 | 2 | -0.31 (0.76) | 0.11 (1.05) |
| auditory_t10 | 2 | -0.19 (0.89) | 0.07 (1.03) |
| auditory_t12 | 22 | -0.72 (1.06) | 0.21 (0.88) |
| **Cognition Composite Scores** | | | |
| Fluid Cognition - Age adjusted | 25 | 0.71 (0.74) | -0.24 (0.96) |
| Crystal Cognition - Age adjusted | 24 | 0.44 (0.80) | -0.15 (1.02) |
| Pattern Comparison Processing Speed - Age adjusted | 1 | 0.43 (0.88) | -0.15 (1.00) |
| **Parental SES Hollingshead-Rendlich** | | | |
| Socio-Economic Status | 2 | 2 | 2 |
| Mother Educational Scale | 7 | 6 | 6 |
| Father Educational Scale | 35 | 6 | 4 |
| **Demographics** | | | |
| Race | 8 | W | W |

## 1.2 Clustering of Mixed Data Types

The cognitive, psychological and demographic data consisted of continuous or numeric and categorical features, hence discrete data (Suppl. Tab. 3), while brain data were continuous. However, mixed continuous and discrete data types create problems for classical clustering algorithms, which work only on a single data type. Clustering algorithms rely on a well-defined distance or similarity measure in a continuous space. This does not apply to categorical data, as there is no inherent ordering or distance measure [1]. Thus, clustering data of mixed types require specifically designed clustering algorithms or mapping of categorical data onto a continuous space [2, 3]. One simple approach is to re-code variables, specifically to one-hot-encode categorical features to binary variables. This is not always the best choice, as it increases the dimensionality of the data, and might influence the performance of clustering algorithms like K-Means [4]. Other approaches that handle mixed data, split the data according to type, and rejoin it after transformation, which is the approach used here. We split our dataset according to numerical and discrete type, and

performed Multiple Correspondence Analysis (MCA) on the categorical data, that maps the categorical data to a continuous space [5], and a Principle Component Analysis (PCA) to the numeric data.

### 1.3 Preprocessing of Behavioural, Psychological and Demographic Data

Preprocessing and data analysis was performed in Python 3.9.7 for behavioural, psychological and demographic data. We used scikit-learn 1.0.2, SciPy 1.7.2 for all analyses, except for MCA, for which we used the package prince 0.7.1.

Prior to our analysis, the data was normalized (see Suppl. Tab. 3 for scaling method of each feature). Scaling of a continuous value x in a feature column X was computed via the mean and standard deviation over the column:

$$\tilde{x} = \frac{x - mean(X)}{std(X)} \ . \tag{1}$$

Skewed features were log-transformed, and age was transformed separately by the following formula:

$$\tilde{x} = \frac{x - min(X)}{max(X) - min(X)} \ . \tag{2}$$

To prepare mixed data for clustering, the data was split according to numerical and categorical features. We performed a PCA on numerical data and an MCA on categorical data [6]. For our analysis, we applied a PCA on numerical data to find the axes of the highest variance and reduce the dimensionality of the data. Results were consistent with other symptom-reduction studies [7]. MCA is an extension of correspondence analysis (CA) for more than two categorical features [8, 9]. It is used on categorical data, similarly as PCA used on numerical data. MCA investigates the association between categorical variables, and, just like a PCA, produces orthogonal components.

As input to our clustering analysis, we combined significant components of the PCA on numerical data and significant components of the MCA. Significance of principal components was determined for components with significantly higher explained variance and inertia for PCA and MCA, respectively, than components of permuted data (i.e., unstructured data; $p<0.05$, 5000 permutations). Components for which explained variance or inertia was less than 5% were discharged. Thus, the data was transformed using PCA and MCA, to allow for combined analysis and reducing dimensionality of the data simultaneously.

### 1.4 Preprocessing of Brain Data

The HCP-EP provides data for various structural and functional Magnetic Resonance Imaging scans for 183 subjects, three subjects had to be excluded with faulty brain scans.

As we are interested in grey matter volume changes we used T1-weighted structural images (N=180, 57 healthy controls, 28 affective psychosis group, 93 non-affective psychosis group and three missing patient specification). During the preprocessing, the structural images of all 180 subjects were segmented into grey matter, white matter, and CSF using Statistical Parametric Mapping, running on MATLAB version 2018b. We used Diffeomorphic Anatomical Registration through Exponentiated Lie Algebra toolbox (DARTEL) [10] to process grey matter images. This procedure creates a sample-specific template representative of all subjects by iteratively aligning all images. Then, the template underwent non-linear registration with modulation for linear and non-linear deformations to the MNI-ICBM152 template. Subsequently, we registered each participant's grey matter map to the group template and smoothed with an 8 $mm^3$ isotropic Gaussian kernel.

Following the preprocessing, we computed an independent component analysis. First, all individually preprocessed grey matter maps were concatenated, creating a 4D file. An absolute grey matter threshold of 0.1 was applied to all images, ensuring that only grey matter voxels were used for the ICA. ICA was performed using the Multivariate Exploratory Linear Optimized Decomposition into Independent Components (MELODIC) method as implemented in the FSL analysis package [11] version 6.0. Data-driven population-based networks of grey matter covariance were derived, performing an ICA on all subjects (n=180). Therefore, this process identifies common spatial components based on the covariation of grey matter patterns across all subjects. We allowed the process to identify 30 components (i.e., structural covariance networks, SCN), as done previously [12–14]. The results were thresholded at z = 3.5 and binarized [15, 16] to eliminate spurious results. Finally, for each participant, grey matter volume was extracted from each of the 30 morphometric networks. Brain regions included in morphometric networks are described in section 4.2 Brain Data.

For two subjects (src_subject_id 4066, 2028), non-brain data contained too many missing values. Thus, all subsequent analyses involving the grey matter volume of the 30 morphometric networks are based on 178 subjects (N=178, 57 healthy controls, 28 affective psychosis group, 93 non-affective psychosis group). For clustering, brain data was corrected for total intacranial volume (TIV), age and sex, using the R stats package, version 4.0.5 [17]. We then performed a PCA on the data, and significant components were used for clustering, applying the same procedures as for non-brain data ($p<0.05$, 5000 permutations).

## 1.5 Preprocessing of Combined Brain and Non-Brain Data

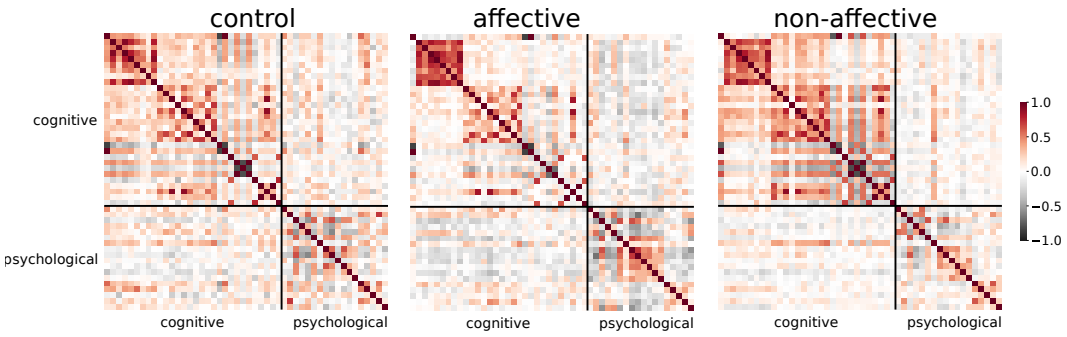
When combining brain and non-brain (i.e., behavioural, demographic, and psychological) data, 151 (46 healthy controls, 24 affective psychosis group, 81 non-affective psychosis group) subjects had matching identifiers and were used for further analysis. The PCA loadings revealed that the PCA performed on all numerical data, brain and non-brain combined, showed almost no differences compared to the PCA conducted on non-brain data only

(Suppl. Fig. 5). The grey matter volumes of networks did not explain any more variance than non-brain data alone when a PCA is performed together across all numerical data. We, therefore, performed a PCA on grey matter values and numerical behavioural data separately and joined significant components of both PCAs with the significant components of the MCA on categorical features (Suppl. Fig. 6, 7) for the combined clustering analysis.

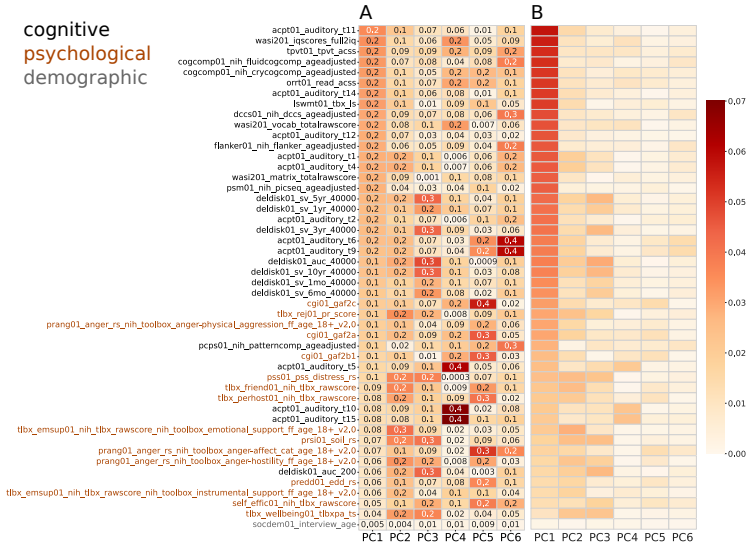
# 2 Results

## 2.1 Non-Brain Data

We identified six significant principal components in numerical features, that captured 58.6% of total variance (permutation test $p<0.05$, 5000 permutations). Cognitive features (i.e., the Auditory Continuous Performance Test, IQ Score, Cognition Composition Score, Reading Recognition Score, and the Picture Vocabulary Test) contributed most to the first principal component, whereas psychological features (e.g. Emotional Support Survey, Perceived Stress Scale or Social Isolation Score) contributed most to the second principal component (Suppl. Fig. 2 for loadings). Visualization of the data in 2D and 3D indicated, that the control group formed a smaller and denser cluster, whereas the non-affective group expanded in the opposite direction of the controls. Affective psychosis subjects were dispersed across controls and non-affective psychosis subjects, which was already implied by the higher inhomogeneity of the affective group. Separation of the three groups was rather apparent on the axis of the first PC, mainly consisting of cognitive features.



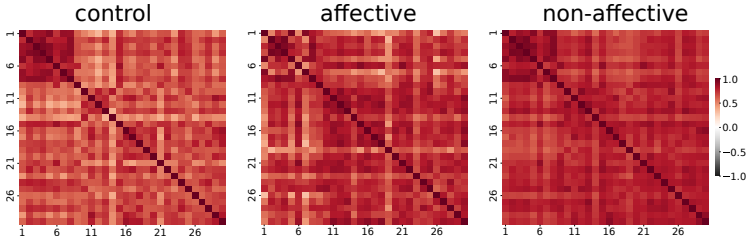

Suppl. Figure 1: **Comparison of behavioural feature correlations between groups.** Correlation between 48 cognitive and psychological features for control (N = 55), affective (N = 51) and non-affective (N = 111) subjects (from left to right).



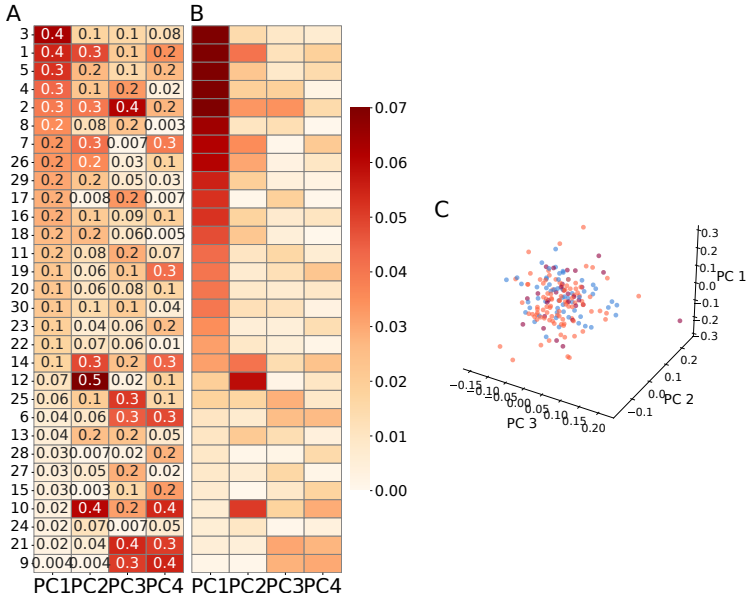

Suppl. Figure 2: **PC loadings of numeric behavioural, psychological and demographic data.** PCA was performed across 217 subjects on 48 numeric behavioural, psychological and demographic features. (**A**) Contribution of each feature to the PCs, in descending order according to the first PC. Features are colour coded according to feature category. Cognitive features contribute most to the first PC, whereas psychometric features contribute mainly to the second PC. (**B**) PC loadings each normalized with variance explained by the PC, thus comparable across components.

## 2.2 Brain Data

We identified four significant components (permutation test $p<0.05$, 5000 permutations) from a PCA that captured 57.1.% of the total variance (Fig. **??**B). Visualization of the data in 2D (Fig. **??**D) and 3D (Supplement, Fig. 4B) did, in contrast to non-brain data, not show a separation of groups. Subjects of all groups were evenly distributed in space. Also non-linear dimensionality reductions did not reveal a separation of groups in 2D or 3D.



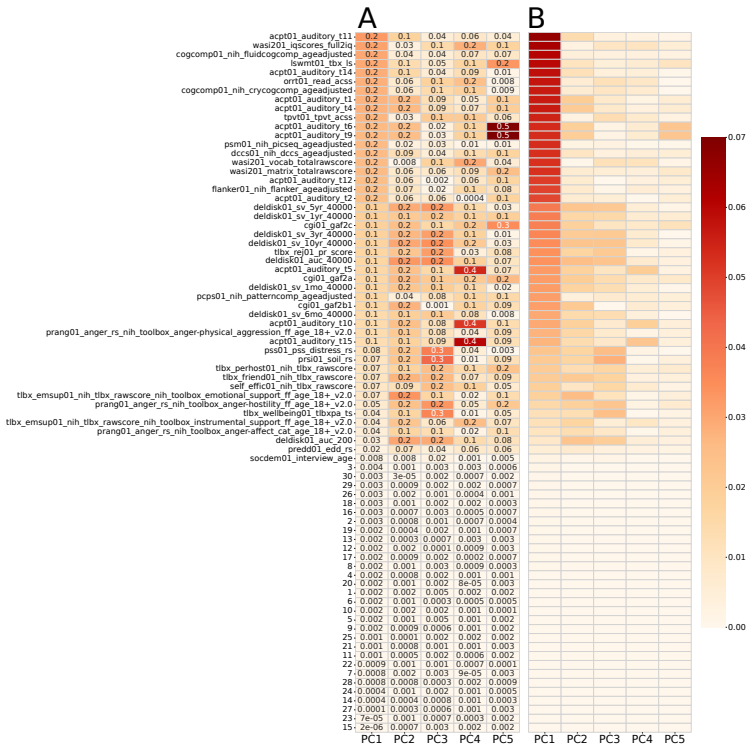

Suppl. Figure 3: **Comparison of brain feature correlations between groups.** Correlation between 30 grey matter values (raw data) for control (N= 57), affective (N= 28) and non-affective (N= 93) subjects (from left to right).



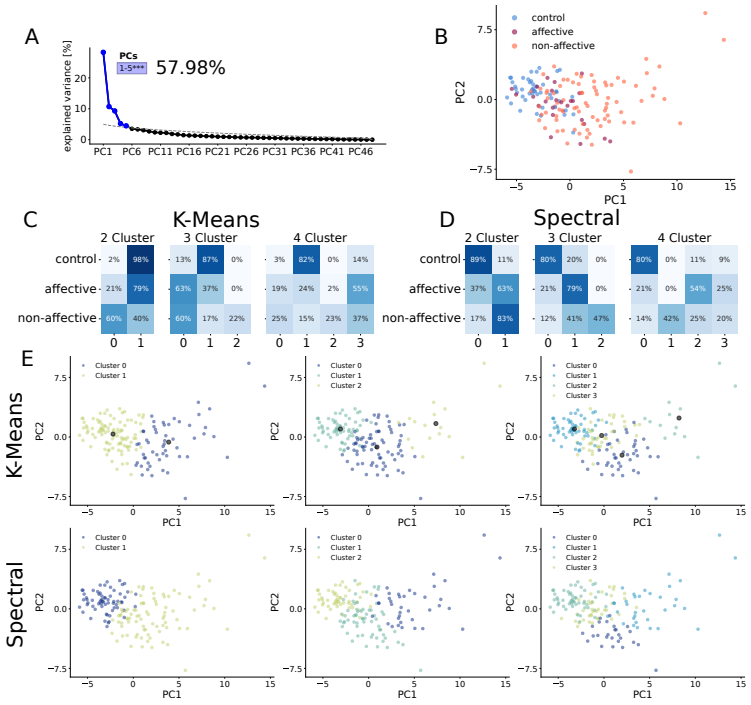

Suppl. Figure 4: **PC loadings and visualization in 3D of brain data.** PCA was performed across 178 subjects on 30 grey matter volumes. (**A**) Contribution of each feature to the PCs, in descending order according to the first PC. (**B**) PC loadings each normalized with variance explained by the PC, making them comparable across components. (**C**) Visualization of brain data in first three PCs, coloured according to subject groups (blue: control, purple: affective, red: non-affective)

## 2.3 Brain and Non-Brain Data

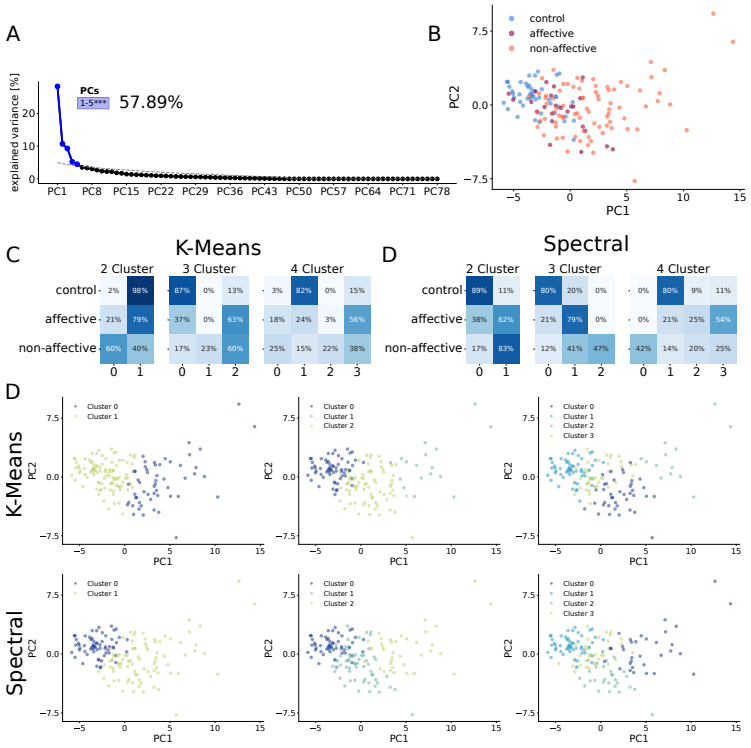
Brain and non-brain data: We identified five significant PCs in behavioural, demographic data, 4 significant PCs in brain data and two significant PCs in categorical data ($p<0.05$, 5000 permutations).




Suppl. Figure 5: **PC loadings of numeric brain and non-brain data.** PCA was performed across 151 subjects on 78 combined features. (**A**) Contribution of each feature to the PCs, in descending order according to the first PC. (**B**) PC loadings each normalized with variance explained by the PC, making them comparable across components.




Suppl. Figure 6: **PCA and clustering analysis on behavioural, psychological and demographic data with 151 subjects used in combined analysis.** (**A**) Variance explained by each of the PCs in % of a PCA performed on all 48 numeric behavioural features across 151 subjects. The first five PCs (blue) survived permutation testing ($p<0.05$, 5000 permutations). Significant components captured 58.0% of all variance. (**B**) The first two PCs are visualized and coloured according to group. (**C**) Percentage of subjects of a group in each cluster, for K-Means clustering analysis with two, three and four clusters on significant PCs. (**D**) Same as C for spectral clustering. (**E**) Clustering result for K-Means (upper) and spectral (lower) analysis visualized on first two PCs for two, three and four clusters.

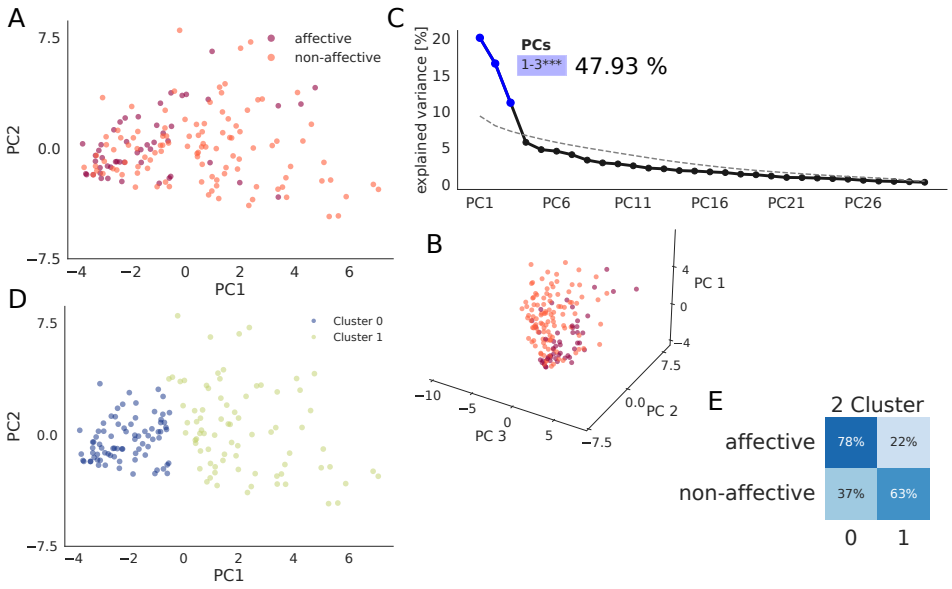



Suppl. Figure 7: **PCA and Clustering analysis on combined brain and non-brain data with 151 subjects.** (**A**) Variance explained by each of the principal components (PCs) in % of a PCA performed on 48 numeric behavioural features and 30 age, sex and TIV corrected gray matter volumes across 151 subjects. The first five PCs (blue) survived permutation testing ($p<0.05$, 5000 permutations). Significant components captured 57.9% of all variance. (**B**) Data visualized on the first two PCs, coloured according to group. There is no apparent change compared to visualization of behavioural data (Fig. 6) because brain data added little to the PCs (Fig. 5) (**C**) Percentage of subjects of a group in each cluster, for K-Means clustering analysis with two, three and four clusters on significant PCs of numeric behavioural data, significant components of brain data and first two components of categorical behavioural data. (**D**) Same as C for spectral clustering. (**E**) Clustering result for K-Means (upper) and spectral (lower) analysis visualized on first two PCs for two, three and four clusters.

## 2.4 Clinical Data

We identified three significant principal components in numerical features, that captured 47.9% of total variance (permutation test $p<0.05$, 5000 permutations). Significant components (PCs 1-3) explained 47.9 % of variance, and patients groups parted mainly on the PC1 axis (Suppl. Fig 8 A-B). K-Means clustering with two clusters on the significant PCs resulted in a cluster containing 63 % of non-affective and 22 % of affective patients and 78% of non-affective and 37 % of affective patients in the other cluster (Supplements, Fig 8 D,

E).

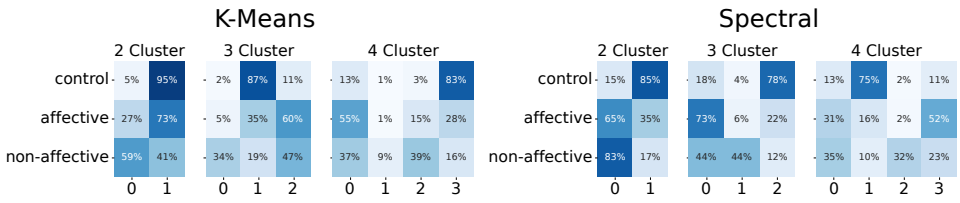



Suppl. Figure 8: **Clustering of clinical scores.** (**A**) The first two PCs are visualized and coloured according to patient group. (**B**) Same as A for first three PCs. (**C**) Variance explained by each of the principal components (PCs) in % of a PCA performed on 30 clinical scores. The first three PCs (blue) survived permutation testing ($p<0.05$, 5000 permutations). Significant components captured 47.9% of all variance. (**D**) K-Means clustering results visualized in the low dimensional space of the PCA. (**E**) Percentage of patients of a group in each cluster, for K-Means clustering analysis with two clusters on significant PCs.

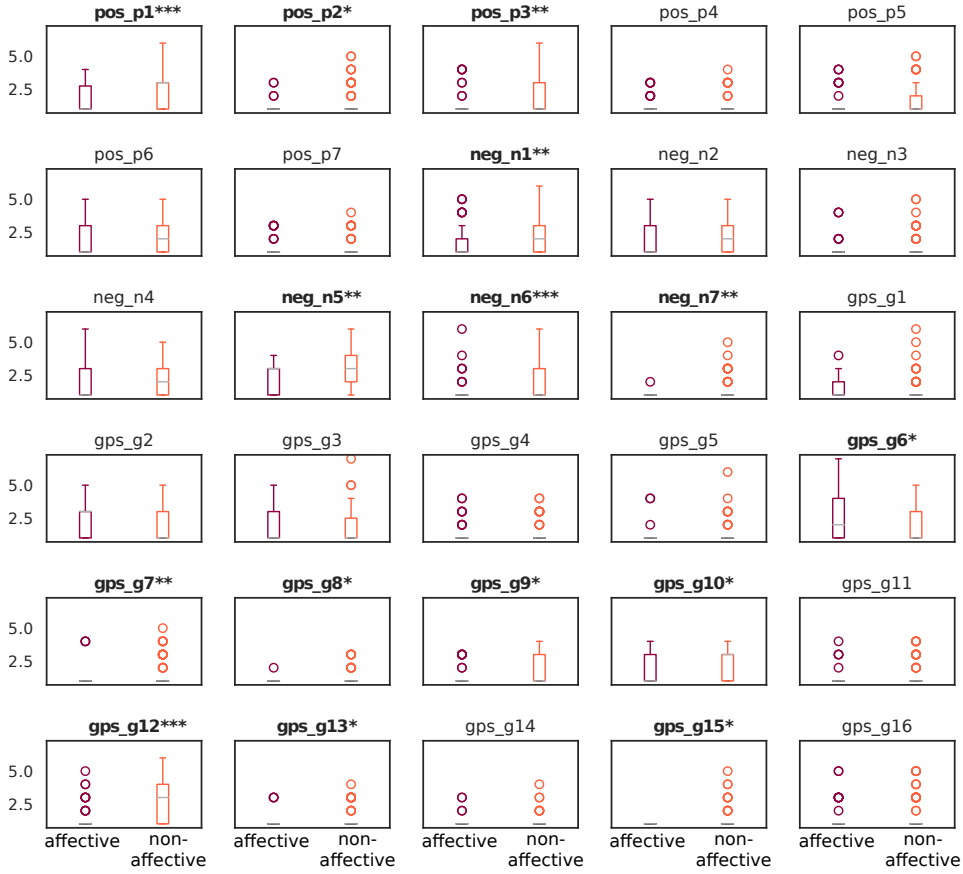
## 2.5 Cluster Exploration

Cognitive scores were compared across patients in specific clusters and control subjects, (Suppl. Fig. 11, see Tab. 2 in main text for statistical comparisons and for boxplots). There were significant group differences in all categories. Differences in cognitive scores supported findings in clinical scores. Patients in cluster 1 showed a similar cognitive performance as controls, as indicated by non-significant differences across all behavioural scores, indicating maintained cognition. Interestingly, cluster 3, which showed only reduced fluid cognition scores compared to cluster 1, showed significant differences compared to controls, with the greatest differences in episodic memory, and selective as well as auditory attention. Cluster 2 showed the worst cognitive performance across all scores (except for impulsivity, for which cluster 0 showed poorer performance), indicating strong cognitive deficits;l. Compared to the cognitively preserved patient group of cluster 1, cluster 0 performed worse across several tasks and scores, including fluid cognition, decision impulsivity, selective attention

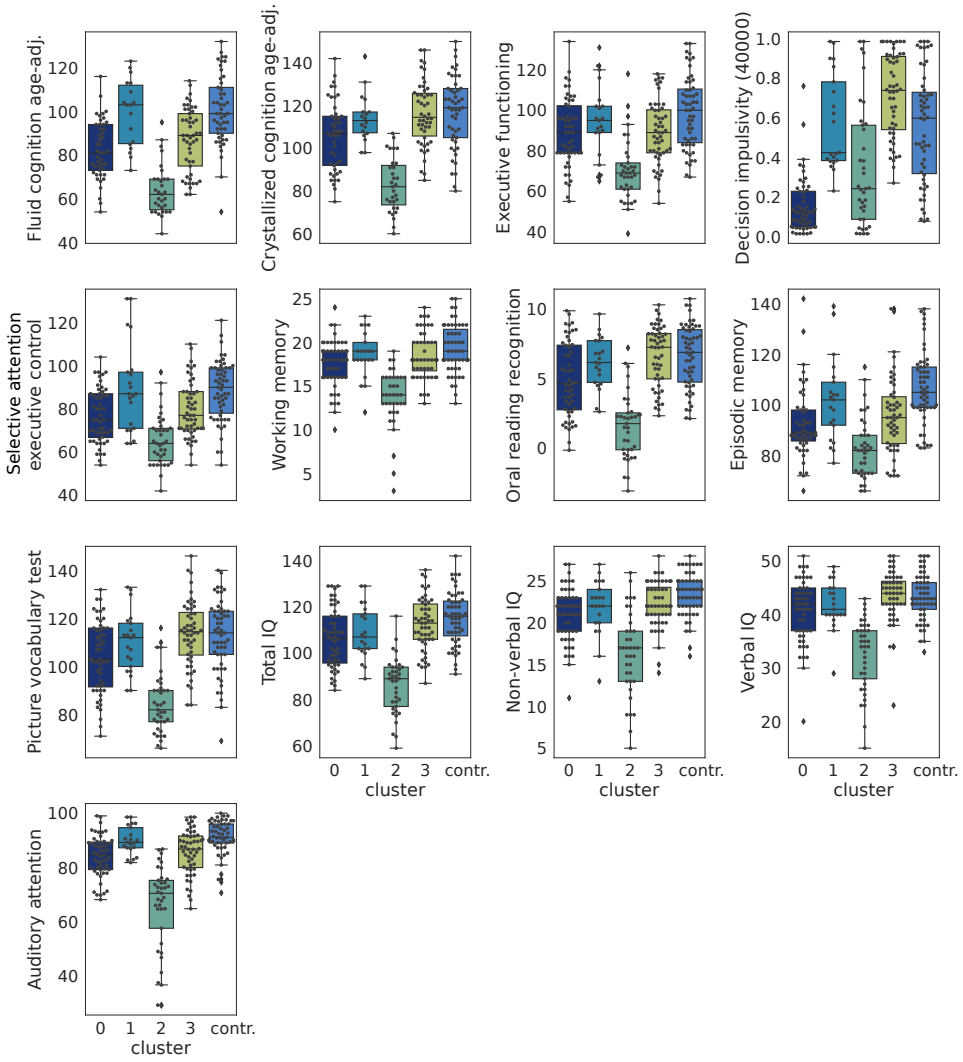
and executive control, episodic memory and auditory attention.




Suppl. Figure 9: **Clustering of Behavioural, Psychological and Demographic Data.** Clustering with significant components of continuous data, without MCA components of categorial data.




Suppl. Figure 10: **Comparison of patient clinical scores.** Individual boxplots represent the distribution of each clinical score, showing data minimum, first quartile, median, third quartile, and data maximum. Outliers are indicated outside the minimum or maximum. Significant scores are indicated with stars, where * corresponds to $p < 0.05$, ** to $p < 0.01$ and *** to $p < 0.001$.




Suppl. Figure 11: **Comparing cognitive features of patients within clusters and control subjects.** Individual boxplots represent the distribution of the performance for each cognitive score, showing data minimum, first quartile, median, third quartile, and data maximum. Individual subjects are overlaid as dots. Outliers are indicated outside the minimum or maximum.

# 3 Extended Discussion

Our correlation analysis revealed that stronger cognitive deficits are linked to stronger negative symptoms, and symptoms or general psychopathology. Although it has been suggested that negative and cognitive symptoms are separable domains which require independent treatment development [18], many studies show an association between cognitive

deficits and negative symptoms across all stages of the disease [19–24]. Our results provide evidence for a link between increased negative symptoms and stronger cognitive deficits across all cognitive domains tested, i.e., crystallized cognition, impulsive decision making, oral reading recognition, picture vocabulary test, total IQ, verbal IQ and auditory attention. The presence of cognitive deficits in combination with increased negative symptoms triggers the question whether one potentially leads, or reinforces, the other. We could argue that deficits in working or episodic memory, or selective attention, which has been reported in risk as well as chronic stages of psychosis, negatively impact social interactions, potentially making them less successful, and leading to poor rapport, one prominent negative symptom, of psychosis. Over time this may lead to social withdrawal, which may in turn negatively impact, e.g. occupational success, having a negative impact on functional outcome. Being excluded from social and/or occupational life may further aggravate cognitive impairment. Further support for this line of thought is provided by the findings that stronger general psychopathology, covering symptoms like depression, disorientation or active social avoidance, assessed using PANSS general psychopathology, are linked to lower fluid and crystallized cognition, worse executive functioning, selective attention and oral reading recognition. As both negative symptoms, including e.g. depression or social withdrawal, and cognitive deficits, have been linked to poorer social functioning [25, 26], identifying individuals with increased risks or increased cognitive dysfunctions may provide new options for patient tailored intervention.

Our correlation analysis revealed that stronger cognitive deficits are linked to stronger negative symptoms, and symptoms or general psychopathology. Although it has been suggested that negative and cognitive symptoms are separable domains which require independent treatment development [18], many studies show an association between cognitive deficits and negative symptoms across all stages of the disease [19–24]. Our results provide evidence for a link between increased negative symptoms and stronger cognitive deficits across all cognitive domains tested, i.e., crystallized cognition, impulsive decision making, oral reading recognition, picture vocabulary test, total IQ, verbal IQ and auditory attention. The presence of cognitive deficits in combination with increased negative symptoms triggers the question whether one potentially leads, or reinforces, the other. We could argue that deficits in working or episodic memory, or selective attention, which has been reported in risk as well as chronic stages of psychosis, negatively impact social interactions, potentially making them less successful, and leading to poor rapport, one prominent negative symptom, of psychosis. Over time this may lead to social withdrawal, which may in turn negatively impact, e.g. occupational success, having a negative impact on functional outcome. Being excluded from social and/or occupational life may further aggravate cognitive impairment. Further support for this line of thought is provided by the findings that stronger general psychopathology, covering symptoms like depression, disorientation

or active social avoidance, assessed using PANSS general psychopathology, are linked to lower fluid and crystallized cognition, worse executive functioning, selective attention and oral reading recognition. As both negative symptoms, including e.g. depression or social withdrawal, and cognitive deficits, have been linked to poorer social functioning [25, 26], identifying individuals with increased risks or increased cognitive dysfunctions may provide new options for patient tailored intervention.

### 3.1 Limitations

This study has several limitations: The ratio of numerical categorical to continuous features is imbalanced. This holds the risk for over-representation of categorical features as they are preprocessed separately; of 6 categorical features two components are generated for clustering, whereas of 54 continuous features 6 PCs are generated. However, in this study, including categorical components did not change the results of our clustering analysis (Suppl. Fig. 9). Generally, K-Means [27] is a commonly used clustering algorithm that performed well on our behavioural data. Nevertheless, there are some drawbacks of this method: First, the algorithm requires a predefined number of clusters. Based on the diagnosis, data contained at least three groups. We used this prior information as basis for the specification of the number of clusters in our analysis. Clustering was performed with two (controls, patients), three (healthy controls, affective psychosis group and non-affective psychosis group) and four clusters (potential subgroups) and the results were compared. Four clusters have been chosen regularly in other clustering work [28, 29]. Identification of five or more clusters require large sample sizes in order to produce reliable and interpretable results. Our sample size does not allow this, because of the restricted number of subjects. Second, K-Means clustering does not work well with non-spherical cluster or clusters with different sizes [30]. We, therefore, also performed spectral clustering, which constructs a similarity graph based on the data (nearest neighbours embedding in our analysis) prior to the clustering and therefore preserving non-linear structure of the data when reducing the dimensions. K-Means clustering is performed on the dimensionality reduced data [31, 32]. This procedure makes the clustering invariant to cluster shapes and densities and by this taking into account clusters of various size and shape, that might arise from the data.

## 4 Features Description

### 4.1 Non-Brain Data

Suppl. Table 2: Of 160 features, 105 were dropped due to missing entries.

| **Dropped Features** | |
|---|---|
| scid-v01-scidd-74 | tpvt01-lavoc-screen |

| Dropped Features | |
|---|---|
| scid-v01-scidd-64 | tpvt01-tlbx-readncorr |
| psychosocial01-psysoc-74 | scid-v01-scidd-66 |
| cgi01-gafrat | socdem01-dem-18b |
| tpvt01-tpvt-fcts | tlbx-wellbeing01-nih-tlbx-fctsc |
| wasi201-iqscores-perfiq | socdem01-sd25h |
| scid-v01-scidd-79 | psychosocial01-psysoc-72 |
| psychosocial01-psysoc-66 | socdem01-sd29 |
| scid-v01-anx-diag | socdem01-cig-smoke |
| socdem01-employcur | socdem01-sd23a |
| socdem01-sd25c | wasi201-ss-blockdesigntscoreperf4 |
| ymrs01-ymrstot | wasi201-sumstscores-total2subtest |
| socdem01-demog-09 | wasi201-iqscores-verbpercentile |
| lswmt01-nih-tlbx-tscore | psychosocial01-psysoc-69 |
| acpt01-auditory-t7 | psychosocial01-psysoc-68 |
| psychosocial01-psysoc-55 | psychosocial01-psysoc-65 |
| socdem01-demo-fam-depression | socdem01-ethnicity |
| socdem01-bio-mother-education | psychosocial01-psysoc-73 |
| wasi201-sumstscores-verbal4subtest | socdem01-sd27c |
| scid-v01-a1d35 | wasi201-iqscores-verbconfintervalto |
| er4001-er40-c-fear | psychosocial01-psysoc-71 |
| psychosocial01-psysoc-67 | flanker01-acc |
| er4001-er40-c-fpn | er4001-er40-c-fpa |
| dccs01-acc | er4001-er40-c-fps |
| er4001-er40-c-noe | socdem01-sd27a |
| pss01-pss-cope-rs | wasi201-iqscores-verbiq |
| socdem01-sd27g | socdem01-smq3 |
| socdem01-fsprels | socdem01-sd25a |
| wasi201-iqscores-perfsumtscores | wasi201-ss-vocabularytscore2 |
| socdem01-sd25g | madrs01-totscr2 |
| socdem01-au1 | socdem01-anthro-weight-calc |
| er4001-er40-c-rtcr | wasi201-sumstscores-total4subtest |
| er4001-er40-c-ang | wasi201-sumstscores-perf4subtest |
| tlbx-rej01-nih-tlbx-fctsc | er4001-er40-c-hap |
| er4001-er40-c-cr | scid-v01-gf-social-scale |
| socdem01-sd24d | socdem01-sd25f |
| tlbx-perhost01-nih-tlbx-fctsc | lswmt01-nih-tlbx-agegencsc |
| socdem01-sd27h | wasi201-ss-similaritiestscoreverbal4 |
| socdem01-psqb20d | socdem01-sd27i |
| flanker01-rt | wasi201-ss-matrixreasoningtscoreperf4 |
| socdem01-psy-health | socdem01-audit12-02a |
| scid-v01-specphob-diag | psychosocial01-psysoc-70 |
| scid-v01-gf-role-scole | socdem01-ps-prime-income |
| wasi201-iqscores-verbconfintervalfrom | socdem01-alcq |
| er4001-er40-c-fpf | psychosocial01-psysoc-75 |
| dccs01-rt | wasi201-ss-vocabularytscoreverbal4 |
| orrt01-read-fcts | socdem01-phq |
| acpt01-auditory-t13 | er4001-er40-c-sad |

| Dropped Features | |
|---|---|
| socdem01-baseline-j-003 | wasi201-iqscores-verbsumtscores |
| acpt01-auditory-t8 | socdem01-deppar |
| acpt01-auditory-t3 | socdem01-sd27d |
| er4001-er40-c-fph | scid-v01-scidd-84 |
| lswmt01-nih-tlbx-fctsc | wasi201-ss-matrixreasoningtscore2 |
| pcps01-nih-tlbx-fctsc | |

Suppl. Table 3: Overview of all features (54) sorted according to cognitive, psychometric and demographic that were used for analysis. Normalization procedure: data was either only scaled or log transformed and scaled. Some features are categorical and were, if necessary, converted into numerical keys.

| Feature | Assessment | Scaling |
|---|---|---|
| **Cognitive** | | |
| acpt01-auditory-t1 | Auditory CPT | logscale |
| acpt01-auditory-t2 | | logscale |
| acpt01-auditory-t4 | | logscale |
| acpt01-auditory-t5 | | logscale |
| acpt01-auditory-t6 | | logscale |
| acpt01-auditory-t9 | | logscale |
| acpt01-auditory-t10 | | logscale |
| acpt01-auditory-t11 | | logscale |
| acpt01-auditory-t12 | | logscale |
| acpt01-auditory-t14 | | logscale |
| acpt01-auditory-t15 | | logscale |
| cogcomp01-nih-fluidcogcomp-ageadjusted | Cognition Composite Scores | logscale |
| cogcomp01-nih-crycogcomp-ageadjusted | | logscale |
| dccs01-nih-dccs-ageadjusted | NIH Toolbox Dimensional Change Card Sort Test | logscale |
| deldisk01-sv-6mo-40000 | Delay Discounting Task | logscale |
| deldisk01-sv-3yr-40000 | | logscale |
| deldisk01-sv-1mo-40000 | | logscale |
| deldisk01-sv-5yr-40000 | | logscale |
| deldisk01-sv-10yr-40000 | | logscale |
| deldisk01-sv-1yr-40000 | | logscale |
| deldisk01-auc-200 | | scale |
| deldisk01-auc-40000 | | scale |
| flanker01-nih-flanker-ageadjusted | NIH Toolbox Flanker Inhibitory Control and Attention Test | logscale |
| lswmt01-tbx-ls | NIH Toolbox List Sorting Working Memory Test | logscale |
| orrt01-read-acss | NIH Toolbox Oral Reading Recognition Test | logscale |
| pcps01-nih-patterncomp-ageadjusted | Pattern Comparison Processing Speed Test | logscale |
| psm01-nih-picseq-ageadjusted | NIH Toolbox Picture Sequence Memory Test | logscale |
| tpvt01-tpvt-acss | NIH Toolbox Picture Vocabulary Test | logscale |
| wasi201-vocab-totalrawscore | Wechsler Abbreviated Intelligence Scale, WASI II | logscale |
| wasi201-matrix-totalrawscore | | logscale |

| Feature | Assessment | Scaling |
|---|---|---|
| wasi201-iqscores-full2iq | | logscale |
| | | |
| **Psychological** | | |
| cgi01-gaf2a | MIRECC GAF Score | logscale |
| cgi01-gaf2b1 | | logscale |
| cgi01-gaf2b2 | | categorial |
| cgi01-gaf2c | | logscale |
| prang01-anger-rs-nih-toolbox-anger | PROMIS Anger | |
| -affect-cat-age-18+-v2.0 | | logscale |
| -physical-aggression-ff-age-18+-v2.0 | | logscale |
| -hostility-ff-age-18+-v2.0 | | logscale |
| predd01-edd-rs | NIH Toolbox Sadness CAT | logscale |
| prsi01-soil-rs | NIH Toolbox Loneliness | logscale |
| pss01-pss-distress-rs | NIH Toolbox Perceived Stress Scale | logscale |
| self-effic01-nih-tlbx-rawscore | NIH Toolbox Self-Efficacy CAT | logscale |
| tlbx-emsup01-nih-tlbx-rawscore-nih-toolbox | NIH Toolbox Emotion Domain | |
| -emotional-support-ff-age-18+-v2.0 | - Emotional Support Survey | logscale |
| -instrumental-support-ff-age-18+-v2.0 | | logscale |
| tlbx-friend01-nih-tlbx-rawscore | - Friendship Survey | logscale |
| tlbx-perhost01-nih-tlbx-rawscore | - Perceived Hostility Survey | logscale |
| tlbx-rej01-pr-score | - Peer Rejection and Perceived Rejection Surveys | logscale |
| tlbx-wellbeing01-tlbxpa-ts | - Psychological Well-Being | logscale |
| | | |
| **Demographic** | | |
| ses01-sestot | Parental SES Hollingshead-Rendlich | categorial |
| ses01-mot-edscale | | categorial |
| ses01-fat-edscale | | categorial |
| socdem01-sex | Demographics Form | categorial |
| socdem01-interview-age | | scaleage |
| socdem01-race | | categorial |

Suppl. Table 4: Number of missing entries per subject

| Subject | Number of missing values |
|---|---|
| NDARBA212KMG | 1 |
| NDARPD149HAG | 1 |
| NDARLG546TEA | 1 |
| NDARGB092FDQ | 1 |
| NDARCN428MJN | 1 |
| NDARXY959VNH | 1 |
| NDARFA038XWN | 1 |
| NDARJG634FY8 | 1 |
| NDARUL529CG9 | 1 |
| NDARJT322WXT | 1 |
| NDARXD502YMF | 1 |
| NDAR-INVLE002YD2 | 2 |
| NDARKW893KM0 | 1 |
| NDARJA321GCN | 1 |
| NDARWV100KAY | 1 |
| NDAR-INVXH347RUY | 1 |
| NDAR-INVWM000VNQ | 1 |
| NDAR-INVAR969RLZ | 1 |
| NDARGY450PB5 | 1 |
| NDARZD427XM6 | 2 |
| NDARNE321KE7 | 1 |
| NDAR-INVEZ580EJ4 | 1 |
| NDAR-INVTV985UYC | 1 |
| NDARMA691MNU | 2 |
| NDARXM887LRX | 1 |
| NDARXM921HKG | 1 |
| NDARWH919JM6 | 1 |
| NDARTB513NAK | 1 |
| NDARKL811MLZ | 1 |
| NDARYP983TGQ | 1 |
| NDARDM382AXR | 1 |
| NDARFV856FT0 | 1 |
| NDARJX175RDU | 1 |
| NDARZF769NFU | 1 |
| NDARXV601GCY | 1 |
| NDAREV475RWF | 1 |
| NDARND142AKJ | 1 |
| NDARJN875AYL | 1 |
| NDARPJ988JAT | 1 |
| NDARKB336YJQ | 1 |
| NDARPZ931MU9 | 1 |
| NDARKA491VAG | 1 |
| NDARAK704TX8 | 2 |
| NDARGR814XTC | 2 |
| NDARWJ204DJA | 3 |
| NDAR-INVFH233LYL | 3 |
| NDAR-INVTU739KL9 | 2 |
| NDARVU141JWN | 2 |
| NDAR-INVPM355ND7 | 2 |
| NDARFZ200EMY | 2 |
| NDARXM919EY1 | 2 |
| NDAR-INVZF057YMK | 2 |
| NDAR-INVGE811UNG | 2 |
| NDARMC681ZFP | 2 |
| NDAR-INVUH339RDB | 2 |
| NDAR-INVKW983VTY | 2 |
| NDARMY791UYA | 2 |
| NDAR-INVWH077PK3 | 2 |
| NDAR-INVEH619UM2 | 4 |
| NDAR-INVMC872XH5 | 2 |
| NDAR-INVXK842PYR | 3 |
| NDAR-INVXF293JK1 | 3 |
| NDAR-INVRD953LXG | 2 |
| NDARNZ015PMM | 2 |
| NDARHL343VKA | 3 |
| NDARJV605VUB | 2 |

## 4.2 Brain Data

**NW1**:

1: 6932, 14, -66, -48, cerebellum VIIIa

2: 6590, -20, -69, -60, cerebellum VIIIa, cerebellum crus

**NW2**:

1: 6763, 8, -48, -28, cerebellum VI

2: 5574, -33, -51, -39, cerebellum VI

**NW3**:

1: 7340, -18, -70, -28, cerebellum VI, cerebellum crus I

**NW4**:

1: 11747, 2, -57, -27, cerebellum I-V

2: 1168, -16, -40, -58, cerebellum VIIIb

3: 134, 21, -42, -56, cerebellum VIIIb

**NW5**:

1: 10338, 6, -74, -38, cerebellum VIIb, cerebellum crus

**NW6**:

1: 5836, -18, -2, -33, parahippocampal gyrus, temporal pole

2: 3948, 34, -6, -51, parahippocampal gyrus, temporal pole

3: 329, -21, -87, -46, cerebellum crus II

**NW7**:

1: 10263, -2, -56, -46, r. cerebellum IX, l. cerebellum IX

2: 489, -27, -56, -36, cerebellum VI, cerebellum V

**NW8**:

1: 6947, -42, -60, -56, cerebellum VIIb, cerebellum crus

2: 5665, 40, -58, -51, cerebellum VIIb, cerebellum crus

3: 352, -14, -6, 10, thalamus

**NW9**:

1: 6039, -40, -4, -38, fusiform, middle temporal gyrus, inferior temporal gyrus

2: 5371, 48, -10, -46, middle temporal gyrus, inferior temporal gyrus

**NW10**:

1: 6981, 3, -78, 8, intracalcarine cortex, lingual gyrus

2: 177, 32, -82, 26 lateral occipital cortex

**NW11**:

1: 12704, 0, 20, 27, cingulate gyrus

2: 176, -33, 33, 33, middle frontal gyrus

3: 109, 34, 27, 30, middle frontal gyrus

4: 100, -8, -4, 15, thalamus

**NW12**:

1: 7408, 2, -64, 18, precuneus, supracalcarine cortex

**NW13**:

1: 5803, -8, -66, 22, precuneus

2: 1015, -28, -62, -33, cerebellum VI, cerebellum crus I

3: 689, 33, -26, 51, postcentral gyrus, precentral gyrus

4: 192, 33, 6, -4, insula

5: 148, 32, -6, -44, fusiform gyrus

6: 112, 36, -36, 39, supramarginal gyrus, superior parietal cortex

**NW14**:

1: 3720, 6, -57, 4, lingual gyrus, precuneus

2: 2335, -14, -48,-9, lingual gyrus, posterior cingulate

**NW15**:

1: 3778, 14, -9, -21, hippocampus, parahippocampal gyrus

2: 3215, -24, -4, -36, hippocampus, parahippocampal gyrus

3: 767, 16, -66, 20, n. accumbens

4: 197, 16, -66, 20, cuneal cortex, supracalcarine cortex

**NW16**:

1: 25138, -4, 48, 3, paracingulate gyrus, cingulate gyrus, middle frontal gyrus, frontal pole

**NW17**:

1: 9529, 0, 28, -14, subcallosal cortex, medial frontal cortex

**NW18**:

1: 8174, -3, 4, -2, putamen, amygdala

2: 7660, -3, 4, -2, putamen, amygdala

**NW19**:

1: 3418, 12, -96, 3, occipital pole

2: 2974, -14, -99, -14, occipital pole

3: 510, 39, -64, -48, cerebellum crus

**NW20**:

1: 3082, -20, 22, 2, caudate, insula

2: 2833, 51, 30, -14, orbitofrontal cortex, insula

**NW21**:

1: 3223, 28, -18, -32, parahippocampal gyrus, fusiform gyrus

2: 2397, -32, -10, -48, fusiform gyrus

3: 339, 48, -52, -18, inferior temporal gyrus

**NW22**:

1: 5026, -12, 10, -18, temporal pole, orbitofrontal cortex

2: 3925, 30, 10, -30, temporal pole, orbitofrontal cortex

3: 2263, 38, -12, -14, insula

4: 1655, -66, -20, -4, superior temporal gyrus, planum temporale

5: 631, -30, -52, -56, cerebellum VIIIa

**NW23**:

1: 2188, 33, -33, 40, postcentral gyrus

2: 1265, -64, -22, 26, postcentral gyrus

3: 172, -50, -27, 10, Heschl gyrus, planum temporale

4: 142, -54, -10, 2, Heschl gyrus, planum temporale

5: 141, -20, -69, -18, cerebellum VI

**NW24**:

1: 6802, -34, -48, -10, inferior temporal gyrus, middle temporal gyrus, fusiform gyrus

2: 4781, 57, -46, -24, inferior temporal gyrus, middle temporal gyrus

3: 406, -38, -58, -57, cerebellum VIIb, cerebellum VIIIa

4: 144, 44, 27, -9, orbitofrontal cortex

5: 32, 26, -24, orbitofrontal cortex, temporal pole

**NW25**:

1: 992, 21, -75, -18, cerebellum VI

2: 316, 22, -84, -38, cerebellum crus

3: 213, 56, -2, -32, middle temporal gyrus

4: 133, 33, -27, -30, fusiform gyrus

5: 126, 40, 10, 38, temporal pole

**NW26**:

1: 10618, -2, -20, 4, thalamus

2: 476, -21, -87, -46, cerebellum crus II

3: 462, -30, -15, 64, precentral gyrus

4: 115, -33, -96, -6, occipital pole

**NW27**:

1: 5041, -4, -48, 24, cingulate gyrus

2: 764, -32, -62, -60, cerebellum VIIIa

3: 188, 12, -74, 22, precuneus

**NW28**:

1: 4057, 51, -15, -18, middle temporal gyrus

2: 1945, -51, -16, -20, middle temporal gyrus

3: 407, 20, -80, -32, cerebellum crus

4: 292, -46, 8, -3, frontal operculum

**NW29**:

1: 8736, -3, 8, 58, paracingulate gyrus, juxtapositional lobule

2: 177, -27, -51, 60, superior parietal lobule

3: 125, 26, -14, 56, precentral gyrus

**NW30**:

1: 2731, -22, 32, 32, superior frontal gyrus, frontal pole

2: 2456, 26, 57, -8, frontal pole

3: 910, 32, -16, -10, putamen

4: 761, -30, -18, -6, putamen

5: 611, 22, -75, -30, cerebellum crus

6: 177, -16, 4, 58, superior frontal gyrus

7: 122, -38, -27, 42, postcentral gyrus

# References


[1] P. Andritsos and P. Tsaparas. "Encyclopedia of Machine Learning and Data Mining". *Encyclopedia of Machine Learning and Data Mining* (2014), pp. 1–6. DOI: 10.1007/978-1-4899-7502-7.

[2] K. Balaji and K. Lavanya. "Clustering Algorithms for Mixed Datasets: A Review". *International Journal of Pure and Applied Mathematics* 118.7 (2018), pp. 547–556.

[3] A. Ahmad and S. S. Khan. "Survey of State-of-the-Art Mixed Data Clustering Algorithms". *IEEE Access* 7.i (2019), pp. 31883–31902. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2903568. arXiv: 1811.04364.

[4] D. T. Dinh, V. N. Huynh, and S. Sriboonchitta. "Clustering mixed numerical and categorical data with missing values". *Information Sciences* 571 (2021), pp. 418–442. ISSN: 00200255. DOI: 10.1016/j.ins.2021.04.076.

[5] L. Phan, H. Liu, and C. Tortora. "K-Means Clustering on Multiple Correspondence Analysis Coordinates". *Researchgate.Net* 1.1 (2019), pp. 1–17. ISSN: 2510-0564. DOI: 10.5445/KSP/1000085952/05.

[6] J. Blasius and M. Greenacre. "Correspondence Analysis and Related Methods in Practice". June 2006, pp. 3–40. DOI: 10.1201/9781420011319.ch1.

[7] J. L. Ji, M. Helmer, C. Fonteneau, J. B. Burt, Z. Tamayo, J. Demšar, B. D. Adkinson, A. Savić, K. H. Preller, F. Moujaes, F. X. Vollenweider, W. J. Martin, G. Repovš, J. D. Murray, and A. Anticevic. "Mapping brain-behavior space relationships along the psychosis spectrum". *eLife* 10 (July 2021). ISSN: 2050084X. DOI: 10.7554/eLife.66968.

[8] M. J. Greenacre. "Theory and Applications of Correspondence Analysis". *The Journal of Animal Ecology* 54.3 (1984), p. 1031. ISSN: 00218790. DOI: 10.2307/4399.

[9] E. J. Beh. "Simple correspondence analysis: A bibliographic review". *International Statistical Review* 72.2 (2004), pp. 257–284. ISSN: 03067734. DOI: 10.1111/j.1751-5823.2004.tb00236.x.

[10] J. Ashburner. "A fast diffeomorphic image registration algorithm". *Neuroimage* 38.1 (2007), pp. 95–113.

[11] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. "Fsl". *Neuroimage* 62.2 (2012), pp. 782–790.

[12] K. Koch, D. Rodriguez-Manrique, O. G. Rus-Oswald, D. A. Gürsel, G. Berberich, M. Kunz, and C. Zimmer. "Homogeneous grey matter patterns in patients with obsessive-compulsive disorder". *NeuroImage: Clinical* 31 (2021), p. 102727.

[13] A. Pichet Binette, J. Gonneaud, J. W. Vogel, R. La Joie, P. Rosa-Neto, D. L. Collins, J. Poirier, J. C. Breitner, S. Villeneuve, E. Vachon-Presseau, et al. "Morphometric network differences in ageing versus Alzheimer's disease dementia". *Brain* 143.2 (2020), pp. 635–649.

[14] Y. Zeighami, M. Ulla, Y. Iturria-Medina, M. Dadar, Y. Zhang, K. M.-H. Larcher, V. Fonov, A. C. Evans, D. L. Collins, and A. Dagher. "Network structure of brain atrophy in de novo Parkinson's disease". *Elife* 4 (2015), e08440.

[15] M. Beckmann, H. Johansen-Berg, and M. F. Rushworth. "Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization". *Journal of Neuroscience* 29.4 (2009), pp. 1175–1190.

[16] F. Knolle, S. S. Arumugham, R. A. Barker, M. W. Chee, A. Justicia, N. Kamble, J. Lee, S. Liu, A. Lenka, S. J. Lewis, et al. "Grey matter morphometric biomarkers for classifying early schizophrenia and PD psychosis: a multicentre study". *medRxiv* (2022).

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013.

[18] P. D. Harvey, D. Koren, A. Reichenberg, and C. R. Bowie. "Negative Symptoms and Cognitive Deficits: What Is the Nature of Their Relationship?" *Schizophrenia Bulletin* 32.2 (Apr. 2006), pp. 250–258. ISSN: 0586-7614. DOI: 10.1093/schbul/sbj011.

[19] L. Leanza, L. Egloff, E. Studerus, C. Andreou, U. Heitz, S. Ittig, K. Beck, M. Uttinger, and A. Riecher-Rössler. "The relationship between negative symptoms and cognitive functioning in patients at clinical high risk for psychosis". *Psychiatry Research* 268 (2018), pp. 21–27.

[20] A. McCleery and K. H. Nuechterlein. "Cognitive impairment in psychotic illness: prevalence, profile of impairment, developmental course, and treatment considerations". *Dialogues in clinical neuroscience* (2022).

[21] C. L. Hovington, M. Bodnar, R. Joober, A. K. Malla, and M. Lepage. "Impairment in verbal memory observed in first episode psychosis patients with persistent negative symptoms". *Schizophrenia research* 147.2-3 (2013), pp. 223–229.

[22] P. M. Grant and A. T. Beck. "Defeatist beliefs as a mediator of cognitive impairment, negative symptoms, and functioning in schizophrenia". *Schizophrenia bulletin* 35.4 (2009), pp. 798–806.

[23] L. Glenthøj, J. R. M. Jepsen, C. Hjorthøj, N. Bak, T. Kristensen, C. Wenneberg, K. Krakauer, M. Nordentoft, and B. Fagerlund. "Negative symptoms mediate the relationship between neurocognition and function in individuals at ultrahigh risk for psychosis". *Acta Psychiatrica Scandinavica* 135.3 (2017), pp. 250–258.

[24] D. S. O'Leary, M. Flaum, M. L. Kesler, L. A. Flashman, S. Arndt, and N. C. Andreasen. "Cognitive correlates of the negative, disorganized, and psychotic symptom dimensions of schizophrenia". *The Journal of neuropsychiatry and clinical neurosciences* 12.1 (2000), pp. 4–15.

[25] K. Kaneko. "Negative Symptoms and Cognitive Impairments in Schizophrenia: Two Key Symptoms Negatively Influencing Social Functioning". eng. *Yonago acta medica* 61.2 (June 2018), pp. 91–102. ISSN: 0513-5710. DOI: 10.33160/yam.2018.06.001.

[26] M. Strassnig, C. Bowie, A. E. Pinkham, D. Penn, E. W. Twamley, T. L. Patterson, and P. Harvey. "Which levels of cognitive impairments and negative symptoms are related to functional deficits in schizophrenia?" *Journal of Psychiatric Research* 104 (2018), pp. 124–129.

[27] J. A. Hartigan and M. A. Wong. "A K-Means Clustering Algorithm". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (Mar. 1979), pp. 100–108. ISSN: 00359254. DOI: 10.2307/2346830.

[28] K. E. Lewandowski, J. T. Baker, J. M. McCarthy, L. A. Norris, and D. Öngür. "Reproducibility of Cognitive Profiles in Psychosis Using Cluster Analysis". *Journal of the International Neuropsychological Society* 24.4 (2018), pp. 382–390. DOI: 10.1017/S1355617717001047.

[29] M. J. Green, L. Girshkin, K. Kremerskothen, O. Watkeys, and Y. Quidé. "A systematic review of studies reporting data-driven cognitive subtypes across the psychosis spectrum". *Neuropsychology Review* 30.4 (2020), pp. 446–460.

[30] L. Ertöz, M. Steinbach, and V. Kumar. "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data". *Proceedings*. Society for Industrial & Applied Mathematics (SIAM), May 2003, pp. 47–58. DOI: 10.1137/1.9781611972733.5.

[31] A. Y. Ng, A. Y. Ng, M. I. Jordan, and Y. Weiss. "On Spectral Clustering: Analysis and an algorithm". *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS* 14 (2001), pp. 849–856.

[32] D. Niu, J. G. Dy, and M. I. Jordan. *Dimensionality Reduction for Spectral Clustering*. Tech. rep. June 2011, pp. 552–560.