Yeung Hon Wah (Orcid ID: 0000-0002-4993-9014)
Stolicyn Aleks (Orcid ID: 0000-0002-1147-7539)
Buchanan Colin (Orcid ID: 0000-0001-6501-628X)
Cox Simon (Orcid ID: 0000-0003-4036-3642)

Spectral Clustering Based on Structural Magnetic Resonance Imaging and its Relationship with Major Depressive Disorder and Cognitive Ability

Hon Wah Yeung [1], Xueyi Shen[1], Aleks Stolicyn[1], Laura De Nooij[1], Mathew A. Harris[1], Liana Romaniuk[1], Colin R. Buchanan[2], Gordon D. Waiter[3], Anca-Larisa Sandu[3], Christopher J. McNeil[3], Alison Murray[3], J. Douglas Steele[4], Archie Campbell[5], David Porteous[5,6], Stephen M. Lawrie[1], Andrew M. McIntosh[1,5], Simon R. Cox[2], Keith M. Smith[7,8], and Heather C. Whalley[1]

[1]Division of Psychiatry, University of Edinburgh, Edinburgh, UK

[2]Lothian Birth Cohorts group, Department of Psychology, University of Edinburgh, Edinburgh, UK

[3]Aberdeen Biomedical Imaging Centre, Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK

[4]School of Medicine, University of Dundee and Department of Neurology, NHS Tayside, Ninewells Hospital and Medical School, UK

[5]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

[6]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK

[7]Usher Institute, University of Edinburgh, Edinburgh, UK

[8]Health Data Research UK, London, UK

Correspondence:

Hon Wah Yeung, Division of Psychiatry, University of Edinburgh, Edinburgh, UK. Email: h.w.yeung@sms.ed.ac.uk

Abstract

There is increasing interest in using data-driven unsupervised methods to identify structural underpinnings of common mental illnesses, including Major Depressive Disorder (MDD) and associated traits such as cognition. However, studies are often limited to severe clinical cases with small sample sizes and most do not include replication.

Here, we examine two relatively large samples with structural magnetic resonance imaging (MRI), measures of lifetime MDD and cognitive variables: Generation Scotland (GS subsample, $N = 980$), and UK Biobank (UKB, $N = 8900$); for discovery and replication, using an exploratory approach. Regional measures of FreeSurfer derived cortical thickness (CT), cortical surface area (CSA), cortical volume (CV) and subcortical volume (subCV) were input into a clustering process, controlling for common covariates. The main analysis steps involved constructing participant K-nearest neighbour graphs, and graph partitioning with Markov Stability to determine optimal clustering of participants. Resultant clusters were i) checked whether they were replicated in an independent cohort, and ii) tested for associations with depression status, and cognitive measures.

Participants separated into two clusters based on structural brain measurements in GS subsample, with large Cohen's d effect sizes between clusters in higher order cortical regions, commonly associated with executive function and decision making. Clustering was replicated in the UKB sample, with high correlations of cluster effect sizes for CT, CSA, CV and

subCV between cohorts across regions. The identified clusters were not significantly different with respect to MDD case-control status in either cohort (GS subsample: $p_{FDR} = 0.2239 - 0.6585$; UKB: $p_{FDR} = 0.2003 - 0.7690$). Significant differences in general cognitive ability were, however, found between the clusters for both datasets; for CSA, CV, subCV (GS subsample: $d = 0.2529 - 0.3490$, $p_{FDR} < 0.005$; UKB: $d = 0.0868 - 0.1070$, $p_{FDR} < 0.005$).

Our results suggest that there are replicable natural groupings of participants based on cortical and subcortical brain measures, which may be related to differences in cognitive performance, but not to the MDD case-control status.

Keywords— Major Depressive Disorder, Cognition, Machine Learning, Clustering, Markov Stability, Structural Neuroimaging

Abbreviations: CHAMP, Convex Hull of Admissible Modularity Partitions; CIDI-SF, Composite International Diagnostic Interview – Short Form; CSA, cortical surface area; CT, cortical thickness; CV, cortical volume; DSM, Diagnostic and Statistical Manual of Mental Disorders; DSy, digit-symbol coding; GS, Generation Scotland; ICV, intracranial volume; k-NN, K-nearest neighbour; KW, Kruskal-Wallis; LM, logical memory; Matrix, matrix reasoning; MDD, Major Depressive Disorder; MHV, Mill Hill Vocabulary; ML, Machine Learning; MRI, magnetic resonance imaging; NMI, normalised mutual information; Pairs Match, Pairs Matching; ProsMemory, Prospective Memory; RT, reaction time; SCID, Structured Clinical Interview for DSM-IV Disorder; subCV, subcortical volume; UKB, UK Biobank; VF, phonetic verbal fluency C-F-L; VI, variation information; VNR, verbal numerical reasoning; WAIS-III[UK], Wechsler Adult Intelligence Scale UK – Third Edition

Introduction

Major Depressive Disorder (MDD) is a heritable and disabling psychiatric condition associated with depressed mood and changes in cognitive function (Malhi and Mann, 2018), resulting in a significant reduction in the quality of life and a substantial burden on the individual, family and society. Many previous studies have reported structural and functional brain alterations associated with depression (Drevets et al., 2008; Jiang et al., 2019). Moreover, psychiatric conditions (including MDD) have been shown to be associated with cognitive alterations (De Nooij et al., 2020). Both psychiatric conditions and cognitive functions are found to have underlying neurobiological mechanisms. With recent advances in brain imaging, computational as well as mathematical techniques, there is increasing interest in developing objective measures that could help classify MDD status and associated traits such as cognition using neuroimaging data. Studies from many different research groups have indicated structural brain differences in MDD using large robust samples. MDD-related cortical thinning was found in orbitalfrontal cortex (Schmaal et al., 2017), medial prefrontal cortex (Treadway et al., 2015), temporal (Zhao et al., 2017), subgenual anterior cingulate cortex (Anderson et al., 2020), lingual gyrus (Suh et al., 2019), precentral (Bos et al., 2018) and par orbitalis (Merz et al., 2018) regions. Some studies also reported lower surface areas in lingual, fusiform, parahippocampal gyrii (Couvy-Duchesne et al., 2018), and subcallosal regions (Wei et al., 2020), as well as cortical volume reduction in prefronal cortex, orbitalfrontal cortex (Grieve et al., 2013), subcallosal regions (Wei et al., 2020), temporal pole, insula lobe (Amidfar et al., 2020) and subgenual anterior cingulate cortex (Niida et al., 2019). While some research indicated MDD-related reduction in thalamus (Schmaal et al., 2016; Webb et al., 2014; Ye et al., 2020), amygdala (Qi et al., 2018), and hippocampus (Nugent et al., 2013), the MDD case-control volumetric differences in subcortical regions have been found to be insignificant in some other studies (Bos et al., 2018; Shen et al., 2017). Furthermore, white matter microstructure (Chen et al., 2017; Shen et al., 2017; van Velzen et al., 2020), functional connectivity (Qiao et al., 2020; Ran et al., 2020)

abnormalities were also found in MDD patients. Although MDD-related brain differences were found in several literature, these studies usually reported small to very small effect sizes. Previous machine learning (ML) studies with structural brain features also show potential for unbiased diagnostic classification (Lebedeva et al., 2017; Patel et al., 2015; Qiu et al., 2014). Features derived from structural magnetic resonance imaging (MRI) have shown promise for MDD case-control classification, with linear or non-linear supported vector machine classifiers achieving accuracies of >70% (Gao et al., 2018). However, the ability of ML to determine case-control status using such features remains uncertain, especially when most existing studies have been conducted on relatively small datasets ($N < 100$), with limited independent replication. In addition, the majority of existing studies focus on clinically ascertained cases and therefore the results may not be generalisable to population or community-based samples (Stolicyn et al., 2020).

While supervised learning methods focus on the core question of whether differences in brain measures characteristic of MDD are sufficient to accurately classify MDD cases from healthy controls, unsupervised learning methods focus on determining whether natural groupings based on brain differences are relevant for MDD. We considered this as a potentially useful approach, since results from unsupervised learning methods, could in turn help us further refine and better understand the disorder. Moreover, clustering has been shown to be an important tool in other areas of medicine, such as in understanding Alzheimer's disease (Alashwal et al., 2019) and different psychiatric disorders (Marquand et al., 2016). Recently other studies have also attempted similar unsupervised clustering analysis approaches on structural (Zhou et al., 2019) and functional imaging data (Drysdale et al., 2017; Tokuda et al., 2018) as a way to identify potential imaging-based data-driven depression subtypes.

In the current study, we applied unsupervised spectral clustering, as an exploratory approach, to data from a relatively large sample of well characterized individuals (MDD cases and controls drawn from a community-based sample, Generation Scotland (Smith et al., 2012), with structural imaging measures, depression phenotyping, and cognitive data). Our rationale was to explore if the effects are observable using unsupervised spectral clustering. Our aim was to identify natural groupings of individuals, characterised by maximally distinct patterns of structural brain properties. We then attempted replication of the clustering in an independent sample with imaging data (UK Biobank (UKB) (Miller et al., 2016)), using regional between-cluster effects as a basis for evaluating replication. Finally, we investigated whether these natural imaging-based groupings are related to distinct clinical and cognitive features of the participants, focussing on those phenotypes that are consistent across cohorts.

Participant graphs were constructed for each FreeSurfer-derived structural metric of cortical thickness (CT), cortical surface area (CSA), cortical volume (CV) and subcortical volume (subCV) subsets separately. Firstly, imaging variables were controlled for age, sex, intracranial volume (ICV) and MRI site, and then normalized. K-nearest neighbour (k-NN) graphs were then constructed based on pairwise distances between each pair of participants, and finally clustering was conducted using a dynamic graph-based Louvain modularity algorithm (Blondel et al., 2008). This was chosen to optimize the Markov Stability (Schaub et al., 2012a) as a measure of the clustering quality instead of the standard modularity measure (Newman, 2006), which has been shown to result in over-partitioning for graphs with strong local structure, such as the k-NN graph (Schaub et al., 2012b). By optimizing Markov Stability, large communities can be revealed at longer Markov times, thus solving the problem of over-partitioning. As such, this method can reveal stable natural groupings within a cohort.

Our main aims were i) to determine whether there was a natural clustering of participants based on structural imaging features, and whether these were replicated in an independent cohort, and, as an exploratory step, ii) to assess whether the clustering results were associated with depression status or cognitive features in both cohorts.

Methods and Materials

Data acquisition and pre-processing

Generation Scotland dataset

Generation Scotland (GS subsample) is a community-based dataset with imaging data, reported previously (Habota et al., 2019; Navrady et al., 2018; Romaniuk et al., 2019; Rupprechter et al., 2020; Smith et al., 2012; Stolicyn et al., 2020). Demographic details of these participants, and for the replication cohort (UKB) are presented in Table 1. Ethical approval for the GS subsample was obtained from the NHS Tayside committee on research (reference 14/SS/0039).

T1 imaging of $N = 1070$ participants from GS subsample, scanned between June 2015 and May 2019, were performed at 2 sites ($N = 544$ from Aberdeen and $N = 526$ from Dundee). Structural measures were derived from T1 images with FreeSurfer version 5.3 (Dale et al., 1999; Fischl et al., 1999, 2004). Mean CT, CSA and CV were derived for 68 cortical regions defined by the Desikan-Killiany atlas (Desikan et al., 2006). Volumes of 21 subcortical structures – including left and right accumbens, amygdala, caudate nucleus, hippocampus, pallidum, putamen, thalamus and four cerebellar regions – were also extracted with FreeSurfer. In total, $N = 980$ participants remained after quality control - removing participants with any missing values, as well as participants whose intracranial volume (ICV) measure and global cortical measures, i.e. overall cortical volume (sum of regional cortical volumes), overall surface area (sum of regional surface areas), were more than three standard deviations away from the sample mean (Stolicyn et al., 2020). Details of MRI acquisition and quality control process are described in Supplementary A.1.2 and A.1.3. Participants whose demographic information was missing were also removed. There were 225 FreeSurfer-derived features available for each participant (204 cortical and 21 subcortical features). Standard Z-score normalization was performed prior to graph construction.

For the GS subsample, there were $N = 980$ participants in total, of whom $N = 302$ were cases with lifetime (current or past) MDD. Diagnosis were established using the Structured Clinical Interview for DSM-IV Disorders (SCID) (First, 1997), and were based on criteria from the Diagnostic and Statistical Manual of Mental Disorders (DSM) (American Psychiatric Association, 2000) (Supplementary A.2.1). Participants were classed as currently depressed if they had an ongoing depressive episode, and as past MDD if they were not depressed at the time of MRI scan but had at least one depressive episode previously (Stolicyn et al., 2020). Participants were classed as recurrent if they had had more than one depressive episode. Data for each participant therefore included MDD status according to the SCID diagnosis described above, single versus recurrent episodes (single: $N = 116$, recurrent $N = 186$).

The cognitive measures were derived from the following tasks: Wechsler Adult Intelligence Scale UK – Third Edition (WAIS-III[UK]) logical memory (LM) part I & II (sum of immediate/delayed recall) (Wechsler et al., 1998), WAIS-III[UK] digit-symbol coding (DSy) (Wechsler), phonemic verbal fluency C-F-L (VF) (Lezak, 1995), Mill Hill Vocabulary (MHV) (Raven et al., 2003) and matrix reasoning tests (Matrix) (Ritchie et al., 1993). Additionally, age, sex, MRI scan site (Aberdeen or Dundee) and Intracranial Volume (ICV) were available and controlled for as described below. Table 1a shows the GS subsample participants characteristics.

UKB dataset

The UK Biobank obtained ethical approval from the NHS Research Ethics Committee (reference11/NW/0382), and our current study was approved by the UKB Access Committee (Project #4844). All participants in both the GS subsample and UK Biobank gave written informed consent.

Data used were the raw T1-weighted volumes were from the second release of UKB MRI data (Jan 2017). All scans were acquired at the same 3T scanner (Siemens Skyra) at one single site (Cheadle). Information on the acquisition parameters can be found in the UK Biobank online Brain Imaging Documentation (https://biobank.ctsu.ox.ac.uk/crystal/docs/ brain_mri.pdf). As with GS subsample, the T1 volumes were processed at the University of Edinburgh with FreeSurfer version 5.3 using default settings, and brain measures were extracted according to the Desikan-Killiany atlas (Desikan et al., 2006). CT, CSA and CV were computed for the 68 cortical regions, alongside volumes of 21 subcortical structures. FreeSurfer parcellations were visually assessed for a variety of errors (Cox et al., 2019b; Stolicyn et al., 2020). Major errors included zero or partial output, substantial skull strip issues or tissue identification errors. Where no major errors were present, parcellations were examined for minor errors including erroneous boundary placement, minor skull stripping issues, and minor tissue omission. Participants with missing values, missing demographic information, as well as those who were outliers in ICV and global cortical measures (as above for GS subsample) were removed, resulting in a dataset with $N = 8900$ participants in total, see Supplementary A.1.4.

Diagnosis of lifetime depression was based on participant responses in the online version of the Composite International Diagnostic Interview - Short Form (CIDI-SF) (Kessler et al., 1998) and made according to the DSM diagnostic criteria (Stolicyn et al., 2020), see Supplementary A.2.2. Data for each participant included lifetime MDD according to the DSM diagnostic criteria.

The cognitive measures were derived from the following tests: Verbal Numerical Reasoning test (VNR, UKB Field ID: 20016.2.0), Reaction Time test (RT, UKB Field ID: 20023.2.0, log-transformed), Pairs Matching test (Pairs Match, UKB Field ID: 399.2.2, log (x+1) transformed) and Prospective Memory test (ProsMemory, UKB Field ID: 20018.2.0) (Fawns-Ritchie and Deary, 2020). While it would have been optimal to match the tasks more closely to those in our GS subsample, Matrix pattern completion and symbol digit substitution tasks were introduced later and therefore were not conducted concurrent with the imaging assessment for the N= 8900 participants in this study. In a recent investigation however, we note that he current 4 cognitive variables that were concurrent with imaging correlated well with other more detailed cognitive tasks within the UKB and with standard validated psychometric indicators of g (Fawns-Ritchie and Deary, 2020). Additionally, age, sex and

Intracranial Volume (ICV) were available and controlled for as described below. Table 1b shows the UKB participants characteristics.

Cognitive function *g*-factor extraction

In addition to the measures from individual tasks, we also derived a measure of general cognitive function - *g*-factor for participants within each cohort (Deary et al., 2010; Johnson and Bouchard Jr, 2005) and assessed the association between the clusters and the derived *g*-factor. The measure of general cognitive ability (*g*-factor) was a well replicated phenomena in psychological sciences (Deary et al., 2010; Warne and Burningham, 2019). Previous research have shown that the *g*-factor derived from entirely different sets of cognitive tests correlated well with each other, given that the set of cognitive tasks covers a sufficiently broad cognitive domain (Johnson et al., 2004, 2008b).

The *g*-factor here is the first factor score from factor analysis employed using the *factoran* function in MATLAB 2020a. For GS subsample, the *g*-factor was based on measures from the matrix reasoning test (Ritchie et al., 1993), verbal fluency test (Lezak, 1995), Mill Hill vocabulary test (Raven et al., 2003), logical memory (Wechsler et al., 1998) and digit-symbol coding tests (Wechsler). Proportion of variance explained by *g*-factor in GS subsample was 26.0%. For UKB, *g*-factor was computed using measures from all the available UKB cognitive tasks stated above using the same process as in GS subsample. Proportion of variance explained by the *g*-factor in UKB was 14.7%. Details of the loadings can be found in Supplementary A.8, Table A.4a and Table A.4b.

Correction for covariates

Correction for covariates was performed by residualizing each brain measure with respect to sex, age, MRI site and ICV using linear regression models (Alfaro-Almagro et al., 2021; Becher, 1992; Dukart et al., 2011; Kostro et al., 2014; More et al., 2021; Snoek et al., 2019). We additionally conducted a Kruskal-Wallis (KW) test to confirm that no group differences remained on the basis of these covariates between identified clusters for both GS subsample and UKB (see Supplementary A.6, Table A.1 and Table A.2).

Graph construction

We applied a dynamical graph community detection approach to assess clustering of participants, which involved graph construction as the first step. Without known graph geometry, the graph was determined by the type of construction and the distance function chosen for the pairwise distance matrix based on the structural variables and the type of construction.

Defining distance between participants

The pairwise distance matrix D is defined as $D_{ij} = d(x_i, x_j)$, where x_i and x_j are vectors of regional measures (CV, CSA, CT or subCV) per participants in the data and $d(\cdot, \cdot)$ is the distance function to be specified. We used the standard Euclidean distance:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \text{ for } x, y \in \mathbb{R}^n \tag{2.1}$$

to determine similarity between participants.

Euclidean distance was chosen as other distance functions typically have more assumptions and constraints on the dataset. For example, cosine dissimilarity is typically used for non-centered and time-varying data which was not the case here. As a preprocessing step, we applied standard Z-score normalisation to all measures before calculating the pair-wise Euclidean distances to avoid bias in features with broad value ranges.

K-nearest neighbour Graph Construction

We constructed k-nearest neighbour (k-NN) graphs from the pairwise-distance matrices computed above. In the k-NN graph, each data point (in this case, participant) is connected to the k closest other data points, as found in the distance matrix, D. It can be formulated as:

$$A_{ij} = \begin{cases} 1 & \text{if } d_{ij} \leq d_i^{(k)} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

where d_{ij} is the direct distance from node i to node j, and $d_i^{(k)}$ is the distance of the k-th nearest neighbour from node i. The resulting graph is binary and undirected. Different values of k were tested in order to determine the optimal model for the respective graph construction by verifying the Markov Stability measure on the networks.

Optimal graph partitioning

We used Markov Stability, instead of modularity, as the objective function for the Louvain algorithm for community detection in our graphs (i.e. optimal clustering) (Schaub et al., 2012a). Contrary to other common community detection methods (e.g k-means clustering, hierarchical clustering or graph modularity optimisation), Markov Stability adopts a dynamics-based framework to uncover community structure. Graph partitions can be ranked and compared at each optimization time-step which helps identify stable, optimal partitions (Delvenne et al., 2013). We briefly describe the Louvain algorithm and the Markov Stability measure below. Liu et al. validated the Markov Stability method on several real datasets by comparing with other popular clustering methods in Liu et al., and it had achieved the best normalised mutual information (NMI) values on average (Liu and Barahona, 2020). Moreover, while the number of clusters are required for initialisation for other clustering methods, this clustering technique can perform the clustering in a completely unsupervised manner.

The Louvain algorithm

The Louvain method is a greedy algorithm for graph community detection which typically optimises modularity of the graph partitions. The modularity is used widely to measure the strength of division of a network into clusters. Details and formulation of modularity is included in Supplementary A.3. In the first phase, each node (data point) is assigned its own group, and hence the clusters are defined by individual nodes. Then, for each node i, we evaluate the modularity increment of removing i from its community and putting it into the community of j. At each step, the movement that leads to the largest increase in modularity is chosen. The algorithm repeats the same process until no further movement of nodes can lead to an increase in modularity (Blondel et al., 2008). At this stage, the local maximum is achieved.

The second phase consists of forming a new network from the communities found during the first phase, i.e. treating the communities in the original graph as nodes in the new network. The sum of weights of edges, w_k, within the same community is represented as a self-loop for that community, while edges between new nodes are defined by the sums of respective weights of inter-community links. This can be interpreted as a coarse version of the original graph. The process in the first phase is then applied to the new network. The two phases are then repeated until modularity is optimized and a hierarchy of communities is produced, and this marks the end of a Louvain run (Blondel et al., 2008).

In our study, we applied the Louvain algorithm with optimisation of Markov Stability measure instead of modularity measure for more optimal community detection. We describe the concept of Markov Stability below.

Markov Stability

Markov Stability is a measure of quality of graph community structure (Delvenne et al., 2010). Although modularity is the default measure of partitioning quality in the Louvain algorithm, optimising modularity can lead to over- or under-partitioning of the graph, and detection of less natural groupings (Fortunato and Barthelemy, 2007; Schaub et al., 2012b). Compared to modularity, optimising Markov Stability takes into account the different time scales within the partitioning algorithm (in our case the Louvain algorithm), with finer communities detected as optimal at earlier partitioning time-steps, and larger communities at later time-steps – which leads to more natural groupings. Markov Stability measure is based on running random walks on the graph and recording which groupings appear most natural for each time scale according to the walk process, with length of each walk determined by the time scale (Delvenne et al., 2013, 2010). Details of the Markov Stability calculation are described in (Delvenne et al., 2010; Lambiotte et al., 2014). Further details on relation of Markov Stability to modularity, as well as how modularity can be replaced with Markov Stability in the Louvain algorithm, can be found in section A.4 in the supplementary material.

Assessing clustering robustness

Since several runs of the Louvain algorithm are needed to define optimal partitions, we completed 100 runs of the Louvain Algorithm for each time-step.

Consistency of graph partitioning at each time-step between different Louvain algorithm runs was measured by the average Variation Information (VI) between all pairs of partitions from different runs, evaluated as:

$$VI(\mathcal{P},\mathcal{P}') = \frac{2H(\mathcal{P},\mathcal{P}') - H(\mathcal{P}) - H(\mathcal{P}')}{H(\mathcal{P},\mathcal{P}')}$$
$$= \frac{H(\mathcal{P},\mathcal{P}') - I(\mathcal{P},\mathcal{P}')}{H(\mathcal{P},\mathcal{P}')} \tag{2.3}$$

with

$$I(P,P') = H(P) + H(P') - H(P,P') \tag{2.4}$$

where $I(P,P')$ is the mutual information, and $H(P)$, $H(P')$ and $H(P,P')$ are Shannon entropies used to measure the amount of information contained in partitioning P. Division by $H(P,P')$ is for normalization. In the following sections, we denoted variational information across different Louvain runs by VI, and denoted variational information across different timesteps by VI(t,t').

For each Louvain run, a different initial condition (i.e. the order of nodes being scanned during each merging step in the first phase) was chosen, so the effect of perturbation on partitioning results could be assessed. We assessed the consistency of partitions at each chosen time-point and persistence of the number of communities over the time scale to choose optimal partitions (Delmotte et al., 2012). When more than one partition was considered as stable over a time scale, the clustering partition that remained stable for the longest time period was selected as the most stable.

Stability postprocessing

Stability post-processing applied in the current study is conceptually similar to the Convex Hull of Admissible Modularity Partitions (CHAMP) method described in Weir et al. (Weir et al., 2017). The Louvain algorithm with stability optimisation was run 100 times with 500 time-steps on each run (the 500 time-steps were logarithmically spaced from 1 to 100), on the k-NN graphs constructed with $k = 5,7,9,11$. For each of the 500 time-steps, an optimal graph partitioning was defined across the entire 100 Louvain runs. As a final post-processing step, the defined optimal graph partitions for each time-step were updated by considering partitions in all other time-steps. Details of the post-processing function can be found in Supplementary A.5.

Assessing clustering consistency

We applied Normalized Mutual Information (NMI) to assess consistency between optimal partitions identified when different k values were used for constructing the graphs. NMI measures the information shared by 2 partitions, C_i and C_j. In other words, it measures to what extent knowing about C_i reduces the uncertainty about C_j. The NMI is defined as (Kvålseth, 2017):

$$NMI(\mathcal{P},\mathcal{P}') = \frac{I(\mathcal{P},\mathcal{P}')}{\sqrt{H(\mathcal{P})H(\mathcal{P}')}} \tag{2.5}$$

where $H(P)$ is again, the Shannon entropy. If the 2 partitions are independent, the NMI is 0. If the 2 partitions are exactly the same, then NMI is equal to 1. Another alternative we proposed is the accuracy measure, which is formulated as:

$$\text{Accuracy} = \frac{\text{number of correctly classified samples}}{\text{number of samples in data}}$$

Assessing reproducibility and relation of clusters to cognition and MDD

After stability optimisation and testing for robustness and consistency, we assessed the reproducibility of the partitioning results and tested associations of clusters from the stable partitions with variables of interest.

To evaluate whether clustering was similar in GS subsample and UKB, we computed Pearson correlations between the cluster effects Cohen's d values in GS subsample and in UKB for regions in each of the four modalities. For computing the Cohen's-d values, we took the values of each FreeSurfer region across participants and then calculated the standardised mean differences between the two participant clusters identified for each modality. Cohen's-d values indicated the level of contribution of each regional measure to the separation between clusters, hence a strong correlation of Cohen's-d values between GS subsample and UKB would indicate that each measure had a similar contribution to the between-cluster separation in both cohorts, i.e. a measure with large Cohen's-d in GS subsample would have large Cohen's-d in UKB and vice versa.

To assess associations with MDD and cognitive tasks, the KW test was used since the variables were not normally distributed. For cognition, we initially tested association with the general cognitive ability (g-factor) and then the individual cognitive tasks separately for both cohorts. For individual tasks, in GS subsample this involved testing associations with LM, DSy, VF, MHV, Matrix tasks and in the UKB association with VNR, RT, Pairs Match and ProsMemory tasks. The Benjamini-Hochberg procedure was used to correct the p-values across the tasks for each cohort (Benjamini and Hochberg, 1995).

Results

Participant clusters based on brain measures

Clustering results in GS subsample

As an illustration of how the optimal partitions were chosen, we took the partitioning of the 5-NN graph with regional surface area features in the GS subsample dataset as an example (Figure 1). Figure 1 illustrates participant partitioning throughout the time scale, after controlling for covariates. It shows that partitions with 5, 3 and 2 clusters, had large plateaus with few VI(t,t') spikes within the plateaus, while partitions with 3 and 2 clusters also

had low variational information (VI) (across algorithm runs). This indicates that the key partitions were those with 3 and 2 clusters. Similar procedures were used to inspect the other k-NN graphs for different modalities.

Notably, the data were consistently partitioned into two clusters across different k-NN graphs (fig. 2). Strong similarity between partitions from different k-NN graphs for CT and CV is illustrated by high NMI (> 0.7) between 2-cluster partitions of the 11-NN and 2-cluster partitions of either 5-NN, 7-NN or 9-NN graphs (Table 2). Although we see slightly decreased NMI for CSA and subCV, high accuracies (CSA: ≥89.6% ; subCV: ≥94.7%) still indicate strong similarity between partitioning results within each of the four modalities. To show that the resulting clusters based on different metrics were not highly dependent on each other, we computed the NMI between the clusters. The NMI between the clusters based on different modalities was presented in Supplementary A.9, Table A.5a. The low NMI among clusters are also consistent with the increasing numbers of studies reporting low correspondence between these modalities, particularly for area and thickness in terms of genetic influences and associated phenotypes (Panizzon et al., 2009; Winkler et al., 2010; Cox et al., 2018; Grasby et al., 2020).

Figure 2 shows the summary of modules merging along the time scale and reaching an equilibrium of two clusters for all four k-NN graphs. The variational information, VI, across Markov time for different k-NN graphs was low for clustering into 2 and 3, which implied that these partitions were stable (Figure 2). Since the plateau at the 2-cluster partition was the largest (i.e. the algorithm stays at the 2-cluster partition for the largest time period and which is consistent across different k-NN settings), we therefore concluded that the 2-cluster partition was the most stable partition to assess for associations with the clinical and cognitive phenotypes.

We computed differences in brain measures between the two clusters identified in the 11-NN graph in Cohen's d effects for GS subsample. The regions with largest Cohen's d for each modality were as follows: CT - right hemisphere (RH) supramarginal (d = 1.662); CSA - left hemisphere (LH) rostral middle frontal (d = 1.387); CV - LH superior frontal (d = 1.461); subCV - RH ventral diencephalic volume (d = 1.762). Overall, regions with large effect sizes included superior, medial and orbitofrontal regions, temporal and parietal cortices, and subcortically in ventral diencephalic volume, as well as thalamus and hippocampus. Full results are reported in Supplementary B.1.1, Table B.1 – B.4.

Replication of clustering results in UKB

Similar to GS subsample, two clusters were identified within each of the feature modalities (CT, CSA, CV and subCV) for UKB data.

The data were again optimally partitioned into two clusters across different k-NN graphs. Strong similarity between the partitions from different k-NN graphs for CT, CSA and CV was found with high NMI (> 0.7) between the 2-cluster partitions of the 5-NN and the 2-cluster partitions of either 7-NN, 9-NN or 11-NN graphs (Table 3). For subCV, we also saw high accuracies (subCV: ≥ 92.9%). The NMI between the clusters based on different modalities was presented in Supplementary A.9, Table A.5b.

Similar to the GS subsample, we computed differences in brain measures between the two clusters identified in the 5-NN graph in Cohen's d effects for UKB. The regions with largest Cohen's d for each modality were as follows, CT: RH inferior parietal (d = 1.536); CSA: LH superior frontal (d = 1.090); CV: LH precuneus (d = 1.058); subCV: RH ventral diencephalic volume (d = 1.416). Cluster-related differences were highly correlated

between GS subsample and UKB datasets: correlation coefficients were 0.9392, 0.9226, 0.9241 and 0.7931 respectively for CT, CSA, CV and subCV modalities, see Figure 3. The top 20 cortical regions and top 10 subcortical regions driving the clusters' separations as well as the corresponding Cohen's d for each of the four clustering analyses are listed in Table 4. These results indicate that the natural groupings of participants, as well as the regional measures which best separate the identified clusters, were similar across the GS subsample and UKB datasets. Among those top regions, there were at least 70% overlap between the two cohorts. The overlapping regions included ventral diencephalic volume, thalamus and hippocampus for subcortical regions, and superior, medial and orbitofrontal regions, as well as parietal regions for cortical metrics.

Details of effect sizes for the clustering results based on all 4 modalities for both cohorts are reported in Supplementary B.1.1, Table B.1 – B.4. All effect sizes were positive, which implies that across all 4 feature modalities regional measures in one cluster were larger compared to the other cluster, independent of sex, age and ICV differences. Figure 4 shows that the effect sizes were positive for all regions when clustering was based on CSAs. Figures for other modalities can be found in Supplementary B.2, Figure B.1 – B.3.

For both the GS subsample and the UKB, those regions contributing the most to cluster separation included lateral orbitofrontal, post central, precentral, precuneus, rostral middle frontal, superior frontal, superior parietal and supramarginal areas in both hemispheres. Large effect sizes were noted for these regions (GS subsample: $d = 0.8682 - 1.662$, UKB: $d = 0.7761 - 1.536$), including some regions where $d > 1.2$, giving confidence in the separation of the clusters (Lakens, 2013; Sullivan and Feinn, 2012). Those regions contributing least to between cluster separation (most consistent across individuals) were the caudal anterior cingulate cortex, entorhinal cortex, frontal pole and temporal pole in both hemispheres, and parahippocampal gyrus in the left hemisphere (GS subsample: $d = 0.2095 - 0.6893$, UKB: $d = 0.2498 - 0.6389$).

Association between clusters and MDD and cognitive variables in GS subsample and UKB

As stated in the method sections, the KW test was used as the test statistics to determine statistical significance of between-cluster differences for MDD and cognitive variables in GS subsample and UKB, and the p-values were FDR corrected.

Association of clusters with MDD in GS subsample

The clusters were found to have no significant association with the presence of an MDD diagnosis in any of the two-cluster results derived from the four different modalities (CT: $p_{FDR} = 0.2239$; CSA: $p_{FDR} = 0.3777$; CV: $p_{FDR} = 0.2295$; subCV: $p_{FDR} = 0.6585$). We also tested whether the clusters were associated with the severity of depression in GS subsample by only including recurrent cases ($N = 186$) in the MDD group, and found that this was also not significant (CT: $p_{FDR} = 0.9353$; CSA: $p_{FDR} = 0.4020$; CV: $p_{FDR} = 0.9184$; subCV: $p_{FDR} = 0.6906$). Information about the effect sizes between clusters as well as the effect sizes between cases and controls is included in Supplementary B.1.2, Table B.5 – B.8.

Association of clusters with cognitive measures in GS subsample

Table 5 shows the corrected p-values and effect sizes of associations between the cognitive tasks and the clustering result for CT, CSA, CV and subCV modalities. The general cognitive ability (g-factor) was found to be significantly associated with the clustering based on CSA ($d = 0.2771$, $p_{FDR} = 8.69e-5$), CV ($d = 0.3490$, $p_{FDR} = 5.27e-6$) and subCV ($d = 0.2529$, $p_{FDR} = 0.0022$) but not CT. As for individual tasks, the Digit Symbol Coding (DSy), and Matrix Reasoning (Matrix) tests were found to be significantly associated with the clustering based on CSA (DSy: $d = 0.3181$, $p_{FDR} = 4.26e-5$, Matrix: $d = 0.3229$, $p_{FDR} = 1.56e-5$), CV (DSy: $d = 0.4161$, $p_{FDR} = 2.08e-8$, Matrix: $d = 0.4450$, $p_{FDR} = 1.81e-10$) and subCVs (DSy: $d = 0.4036$, $p_{FDR} = 4.19e-7$, Matrix: $d = 0.3057$, $p_{FDR} = 3.88e-4$). In addition, those two tasks were found to be significantly associated with the clustering based on CT (DSy: $d = 0.3455$, $p_{FDR} = 1.68e-5$, Matrix: $d = 0.1971$, $p_{FDR} = 0.0051$). Significant positive effect sizes in FreeSurfer measures (for all of CT, CSA, CV and subCV) were related to positive effect sizes in cognitive measures, and these results were independent of sex, age and ICV differences. These results suggest that participant clusters defined by larger imaging measures may be characterised by better cognitive performance.

Associations of clusters with MDD status and cognitive measures in UKB

As in the GS subsample, clusters in UKB were found to have no significant associations with lifetime MDD diagnosis (CT: $p_{FDR} = 0.7690$; CSA: $p_{FDR} = 0.3059$; CV: $p_{FDR} = 0.2003$; subCV: $p_{FDR} = 0.6703$). The absence of a direct one-to-one correspondence between the cognitive tasks in GS subsample and the UKB precluded a direct replication of the test-specific cognitive findings in GS subsample using UKB data. However, due to the advantages of cross-battery stability conferred by computing a g-factor (Fawns-Ritchie and Deary, 2020; Johnson et al., 2004, 2008b), we replicated the group differences in g-factor. The g-factor were found to be significantly associated with clustering based on CSA ($d = 0.0868$, $p_{FDR} = 2.05e-4$), CV ($d = 0.1070$, $p_{FDR} = 1.24e-5$), subCV ($d = 0.0919$, $p_{FDR} = 4.48e-5$), and also for CT ($d = 0.0573$, $p_{FDR} = 0.0318$). As for individual tasks, the score for Verbal Numerical Reasoning (UKBID:20016.2.0, Fluid Intelligence) was also found to be significantly associated with clustering based on CSA ($d = 0.0999$, $p_{FDR} = 5.19e-5$), CV ($d = 0.1091$, $p_{FDR} = 1.17e-5$), subCV ($d = 0.0923$, $p_{FDR} = 4.48e-5$) and CT ($d = 0.0502$, $p_{FDR} = 0.0318$). There are also significant associations of clusters based on subCV with reaction time ($d = -0.0716$, $p_{FDR} = 0.0011$), and clusters based on CV and CT with prospective memory (CV: $d = 0.0483$, $p_{FDR} = 0.0468$, CT: $d = 0.0485$, $p_{FDR} = 0.0489$) (see Table 6).

Discussion

Summary of results

Overall Summary

In the current study, we employed an exploratory approach and performed unsupervised spectral clustering with k-NN graphs, which were based on pairwise distances in structural brain measures derived with FreeSurfer. We aimed to determine the presence of natural groupings of participants and their relation to lifetime MDD and cognitive ability. The results identified a natural split of the data into two main clusters for each of the four modalities studied, where clustering results for separate modalities were independent of each other. We replicated the natural groupings of participants into two main clusters in each modality in an independent dataset (UKB) based on the highly correlated cluster-related differences between the two cohorts, with correlation coefficients 0.9392, 0.9226, 0.9241 and 0.7931 respectively for CT, CSA, CV and subCV modalities. Moreover, the results were not driven by common covariates, namely, sex, age, MRI site and ICV. It was found that the strongest contributors to the cluster separation were the ventral diencephalic volume, thalamus and hippocampus for subcortical regions ($d = 1.0891 - 1.7621$) and superior, medial and orbitofrontal regions, along with temporal and parietal regions for cortical metrics ($d = 0.8192 - 1.6617$).

The clusters identified were not related to lifetime MDD status in either data set. We also did not find associations with the more severe MDD cases by taking only those with recurrent MDD in the GS subsample (see Sections 3.2.1 and 3.2.3). While we found no relationship with MDD, there was however significant relationships with cognition base on the general cognitive ability (g-factor) in both GS subsample and UKB (Johnson et al., 2008a). The clusters also showed significant relationships with some other specific tests mainly in the domains of reasoning (Matrix Reasoning in GS subsample and Verbal Numerical Reasoning in UKB) and processing speed (WAIS-III^{UK} Digit Symbol Coding score in GS subsample and Reaction Time Task in UKB). Results suggest that the participant clusters defined by larger FreeSurfer measures are in general characterised by better cognitive performance. Apart from MDD status and cognitive abilities, assessing associations of clusters with other variables (for example brain age, stress and social economic status) could also be an interesting future research direction.

A key feature of our work is the use of covariates to ensure that the clusters are not driven by important factors such as age, head size and sex. Prior to our study, Zhou et al. employed supervised feature selection on $N = 3,297$ brain morphometric measures that approximately represented the 3D neuroanatomical integrity of the participants' brains in the UKB as well as $N = 4,316$ demographic, clinical, biological specimen, imaging, genomic, and questionnaire variables for $N = 9,914$ subjects (Zhou et al., 2019). Although using different clustering methods (k-means clustering and hierarchical clustering), Zhou et al. also carried out clustering analysis on all derived neuroimaging measures and also obtained two clusters, of which one cluster showed larger values in all of their top 20 neuroimaging variables. Contrary to the results of our study, their resulting clusters did show differences regarding in mental health variables, including depressive symptoms. However, they did not adjust for basic covariates and found a significant association of clusters with sex, so that the significant between-cluster

differences found in mental health variables, including depressive symptoms, might be driven by the significant sex disparity between the clusters. That the current study and Zhou et al.'s study show mixed results regarding associations with mental health variable may therefore be related to different methodological approaches.

Interpretation of between-cluster effect sizes

The calculated effect sizes, i.e. the Cohen's *d* coefficients, represent the degree of separation between the individuals in the two clusters for each brain region. Most regions had medium to high effect sizes, which indicates that the two clusters were clearly separated (Lakens, 2013; Sullivan and Feinn, 2012).

The greatest effect sizes were seen for CT in RH supramarginal area in GS subsample ($d = 1.662$) and for CT in RH inferior parietal area in UKB ($d = 1.536$) in cortical measures as well as right ventral diencephalic volume for both GS subsample ($d = 1.762$) and UKB ($d = 1.416$) in subcortical measures. The top regions between GS subsample and UKB had a high percentage of overlap, as shown in Table 4. Details of between-cluster effects can be found in Supplementary B.1.1, Tables B.1 – B.4.

In general, for both cohorts, as well as across different cortical metrics, we note that those regions with larger effect sizes tended to be those which are commonly associated with higher cognitive functions such as executive function and decision making (e.g. precuneus, rostral middle frontal gyrus, superior frontal gyrus, lateral orbitofrontal gyrus and superior temporal gyrus) (Barbey et al., 2012; Camilleri et al., 2018). These results are in line with prior work on brain regional correlates of intelligence (Cox et al., 2019a, 2018), and add further reference – via an unsupervised clustering method – that these higher-order cortical regions are related to cognitive ability beyond influences of gross head size, age and sex.

We note that, Cox et al. found that the frontal pole contributed the most to intelligence (Cox et al., 2019a), while in the current study the frontal pole was found to have one of the smallest between-cluster effect sizes. We consider that this difference likely originates from substantial differences in software, anatomical labelling, and analysis methods. Cox et al. employed the UKB-processed FSL FIRST and FSL FAST parcellations, whereas this study used FreeSurfer derived metrics. This is important since there is currently no consensus regarding the definitions of the posterior extent of the frontal pole from structural neuroimaging data and both of these methods uses different atlas definitions (see Bohland et al., 2009; Cox et al., 2014). Further, Cox et al. implemented structural equation models (SEM) targeting the associations between individual ROIs and the g-factor directly, while clustering analysis as used here in general does not directly test associations between ROI measures and other variables. Moreover, the clustering analysis in this study was methodologically driven by structural brain measures and not cognitive ability measures.

Limitations

A non-hypothesis driven graph clustering analysis can generally help to discover population subtypes based on non-linear relationships between independent variables. For clinical studies, the drawback is that the identified subtypes (clusters) are not guaranteed to be clinically relevant. In our case, we found that the partitioning results were not associated with MDD status as measured in the current samples. However, we note that our samples were

relatively healthy, with few individuals having current depression (most cases met criteria for lifetime MDD rather than current MDD). It is possible that using these diagnostic criteria may have contributed to lack of association of clustering results with MDD status. Previous studies of MDD case-control classification, where high accuracies (i.e. ≥ 85 %) were achieved, typically had sample size smaller than 100 and involved clinically ascertained current MDD cases, in some cases with severe or treatment-resistant depression (Johnston et al., 2015; Mwangi et al., 2012; Patel et al., 2015).

In terms of the lack of MDD-related differences between the clusters in the context of previous supervised studies, we cannot exclude the possibility that this may be due to sample differences (our samples are relatively well community-based samples), or that our unsupervised method may not be sensitive enough in its current form to detect these brain features of typically small to very small effect sizes.

Moreover, we also performed clustering analysis without controlling for any covariates as an initial testing of our method. We found that the resulting clusters were associated with MDD status based on CT, CSA and CV measures for GS subsample and CT, CSA, CV subCV measures for UK Biobank (see Supplementary A.7, Table A.3a and Table A.3b) but they were also strongly driven by sex, age, ICV and scan site. Since we were not specifically interested in sex, age or site effects, these regressed out of the brain measures prior to the main analysis. We note however, that without residualisation we do indeed see the expected clustering related to these characteristics. We cannot exclude the possibility however that residualisation may have removed some effects related to MDD.

In spite of the relative invariance of g to cognitive test battery content, we note that sufficient breadth of cognitive domains is an important consideration in deriving a comparable g-factor. For example, it might be considered that GS tests were more nonverbal and fluid when compared to the verbal and crystallised abilities in UKB. Thus, although there were verbal and crystallised elements in the UKB VNR test, that the UKB tests used here were subsequently shown to be relatively good g measures, it is possible that the g measures extracted across the two cohorts showed imperfect correspondence. Nevertheless, our finding that the natural clustering showed g-differences in both cohorts further militates against this as a substantial confounder of our results.

The current study involved using information from MRI scans based on FreeSurfer parcellations according to the Desikan-Killiany atlas. We note that greater information in the form of raw voxel-wise data may be a better representation of the brain structure and may improve the clustering quality. This would, however, significantly increase the computational cost. Future research could also apply clustering analysis on functional MRI data.

Conclusion

We employed a novel unsupervised clustering algorithm to find natural participant groupings within two large independent datasets of brain structural measures. A natural grouping of two clusters was identified in the first dataset (GS subsample) for each of the four studied modalities, and was replicated in the second dataset (UKB). The main regions driving cluster separation were ventral diencephalic volume, thalamus and hippocampus, superior, medial and orbitofrontal regions, along with temporal and parietal regions in both GS subsample and UKB datasets. Although the clusters were not related to lifetime MDD, they were found to be associated with

general cognitive ability (*g*-factor, computed based on multiple cognitive tasks) in both cohorts, and also with specific reasoning tasks, namely the Matrix Reasoning and Digit Symbol Coding tasks in GS subsample and Fluid Intelligence Score in UKB. Regions with relatively high cluster-related effect sizes were the higher order cortical regions, commonly associated with executive function and decision making. Future work could focus both on development and application of machine learning methods to voxel-wise and multimodal brain imaging data as well as looking at associations with other clinically-relevant metrics.

Acknowledgement

Conflict of Interest

AMM has received research support from Eli Lilly, Janssen, and The Sackler Trust. AMM has also received speaker fees from Illumina and Janssen.

Data and Code Availability

The data collected in the STRADL study have been incorporated in the larger Generation Scotland dataset. Non-identifiable information from the Generation Scotland cohort is available to researchers in the United Kingdom and to international collaborators through application to the Generation Scotland Access Committee (access@generationscotland.org) and through the Edinburgh Data Vault (https://doi.org/10.7488/8f68f1ae-0329-4b73-b189-c7288ea844d7). Generation Scotland operates a managed data access process including an online application form, and proposals are reviewed by the Generation Scotland Access Committee.

Data from the UK Biobank resource is available for health-related research upon registration and application through the UK Biobank Access Management System (https://www.ukbiobank.ac.uk/register-apply/).

The code for Markov Stability algorithm for the analysis of undirected weighted graphs can be found on this public repository: https://github.com/michaelschaub/PartitionStability.

ORCID

Hon Wah Yeung https://orcid.org/0000-0002-4993-9014

References

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., and Moustafa, A. A. (2019). The application of unsupervised clustering methods to Alzheimer's Disease. *Frontiers in computational neuroscience*, 13:31.

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L., Bastiani, M., Miller, K. L., Nichols, T. E., and Smith, S. M. (2021). Confound modelling in uk biobank brain imaging. *NeuroImage*, 224:117002.

Amidfar, M., Quevedo, J., Z. Réus, G., and Kim, Y.-K. (2020). Grey matter volume abnormalities in the first depressive episode of medication-naïve adult individuals: a systematic review of voxel based morphometric studies. *International Journal of Psychiatry in Clinical Practice*, pages 1–14.

Anderson, K. M., Collins, M. A., Kong, R., Fang, K., Li, J., He, T., Chekroud, A. M., Yeo, B. T., and Holmes, A. J. (2020). Convergent molecular, cellular, and cortical neuroimaging signatures of major depressive disorder. *Proceedings of the National Academy of Sciences*, 117(40):25138–25149.

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR)* (4th ed., Vol. 1). Arlington, VA: American Psychiatric Association.

Barbey, A. K., Colom, R., Solomon, J., Krueger, F., Forbes, C., and Grafman, J. (2012). An integrative architecture for general intelligence and executive function revealed by lesion mapping. *Brain*, 135(4):1154–1164.

Becher, H. (1992). The concept of residual confounding in regression models and some applications. *Statistics in Medicine*, 11(13):1747–1758.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bohland, J. W., Bokil, H., Allen, C. B., and Mitra, P. P. (2009). The brain atlas concordance problem: quantitative comparison of anatomical parcellations. *PloS one*, 4(9):e7200.

Bos, M. G., Peters, S., van de Kamp, F. C., Crone, E. A., and Tamnes, C. K. (2018). Emerging depression in adolescence coincides with accelerated frontal cortical thinning. *Journal of Child Psychology and Psychiatry*, 59(9):994–1002.

Camilleri, J., Müller, V. I., Fox, P., Laird, A. R., Hoffstaedter, F., Kalenscher, T., and Eickhoff, S. B. (2018). Definition and characterization of an extended multiple-demand network. *NeuroImage*, 165:138–147.

Chen, G., Guo, Y., Zhu, H., Kuang, W., Bi, F., Ai, H., Gu, Z., Huang, X., Lui, S., and Gong, Q. (2017). Intrinsic disruption of white matter microarchitecture in first-episode, drug-naive major depressive disorder: A voxel-based meta-analysis of diffusion tensor imaging. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 76:179–187.

Couvy-Duchesne, B., Strike, L. T., de Zubicaray, G. I., McMahon, K. L., Thompson, P. M., Hickie, I. B., Martin, N. G., and Wright, M. J. (2018). Lingual Gyrus surface area is associated with anxiety-depression severity in young adults: a genetic clustering approach. *Eneuro*, 5(1).

Cox, S., Ritchie, S., Fawns-Ritchie, C., Tucker-Drob, E., and Deary, I. (2019a). Structural brain imaging correlates of general intelligence in uk biobank. *Intelligence*, 76:101376.

Cox, S. R., Bastin, M. E., Ritchie, S. J., Dickie, D. A., Liewald, D. C., Maniega, S. M., Redmond, P., Royle, N. A., Pattie, A., Hernández, M. V., et al. (2018). Brain cortical characteristics of lifetime cognitive ageing. *Brain Structure and Function*, 223(1):509–518.

Cox, S. R., Ferguson, K. J., Royle, N. A., Shenkin, S. D., MacPherson, S. E., MacLullich, A. M., Deary, I. J., and Wardlaw, J. M. (2014). A systematic review of brain frontal lobe parcellation techniques in magnetic resonance imaging. *Brain structure and function*, 219(1):1–22.

Cox, S. R., Lyall, D. M., Ritchie, S. J., Bastin, M. E., Harris, M. A., Buchanan, C. R., Fawns-Ritchie, C., Barbu, M. C., De Nooij, L., Reus, L. M., et al. (2019b). Associations between vascular risk factors and brain MRI indices in UK Biobank. *European heart journal*, 40(28):2290–2300.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, 9(2):179 – 194.

De Nooij, L., Harris, M. A., Adams, M. J., Clarke, T.-K., Shen, X., Cox, S. R., McIntosh, A. M., and Whalley, H. C. (2020). Cognitive functioning and lifetime major depressive disorder in UK Biobank. *European Psychiatry*, 63(1).

Deary, I. J., Penke, L., and Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature reviews neuroscience*, 11(3):201–211.

Delmotte, A., Tate, E. W., Yaliraki, S. N., and Barahona, M. (2012). Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction.

Delvenne, J.-C., Schaub, M. T., Yaliraki, S. N., and Barahona, M. (2013). *The Stability of a Graph Partition: A Dynamics-Based Framework for Community Detection*, volume 2 of *Dynamics On and Of Complex Networks*, pages 221–242. Springer.

Delvenne, J.-C., Yaliraki, S. N., and Barahona, M. (2010). Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29):12755.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968 – 980.

Drevets, W., Price, J., and Furey, M. (2008). Brain structural and functional abnormalities in mood disorders: implications for neurocircuitry models of depression. *Brain Structure and Function*, 213(1-2):93–118.

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D. J., Etkin, A., et al. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23(1):28–38.

Dukart, J., Schroeter, M. L., Mueller, K., Initiative, A. D. N., et al. (2011). Age correction in dementia–matching to a healthy brain. *PloS one*, 6(7):e22193.

Fawns-Ritchie, C. and Deary, I. J. (2020). Reliability and validity of the UK Biobank cognitive tests. *PloS one*, 15(4):e0231627.

First, M. B. (1997). Structured clinical interview for DSM-IV axis I disorders. *Biometrics Research Department*.

Fischl, B., Sereno, M. I., and Dale, A. M. (1999). Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System. *NeuroImage*, 9(2):195 – 207.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., and Dale, A. M. (2004). Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex*, 14(1):11–22.

Fortunato, S. and Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the national academy of sciences*, 104(1):36–41.

Gao, S., Calhoun, V. D., and Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, 24(11):1037–1052.

Grasby, K. L., Jahanshad, N., Painter, J. N., Colodro-Conde, L., Bralten, J., Hibar, D. P., Lind, P. A., Pizzagalli, F., Ching, C. R.,McMahon, M. A. B., et al. (2020). The genetic architecture of the human cerebral cortex. *Science*, 367(6484).

Grieve, S. M., Korgaonkar, M. S., Koslow, S. H., Gordon, E., and Williams, L. M. (2013). Widespread reductions in gray matter volume in depression. *NeuroImage: Clinical*, 3:332–339.

Habota, T., Sandu, A.-L., Waiter, G. D., McNeil, C. J., Steele, J. D., Macfarlane, J. A., Whalley, H. C., Valentine, R., Younie, D., Crouch, N., et al. (2019). Cohort profile for the stratifying resilience and depression longitudinally (stradl) study: A depression-focused investigation of generation scotland, using detailed clinical, cognitive, and neuroimaging assessments. *Wellcome open research*, 4(185):185.

Jiang, X., Shen, Y., Yao, J., Zhang, L., Xu, L., Feng, R., Cai, L., Liu, J., Chen, W., and Wang, J. (2019). Connectome analysis of functional and structural hemispheric brain networks in major depressive disorder. *Translational Psychiatry*, 9(1):136.

Johnson, W. and Bouchard Jr, T. J. (2005). Constructive replication of the visual–perceptual-image rotation model in Thurstone's (1941) battery of 60 tests of mental ability. *Intelligence*, 33(4):417–430.

Johnson, W., Bouchard Jr, T. J., Krueger, R. F., McGue, M., and Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence*, 32(1):95–107.

Johnson, W., Carothers, A., and Deary, I. J. (2008a). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on psychological science*, 3(6):518–531.

Johnson, W., te Nijenhuis, J., and Bouchard Jr, T. J. (2008b). Still just 1 g: Consistent results from five test batteries. *Intelligence*, 36(1):81–95.

Johnston, B. A., Steele, J. D., Tolomeo, S., Christmas, D., and Matthews, K. (2015). Structural MRI-based predictions in patients with treatment-refractory depression (TRD). *PloS one*, 10(7):e0132958.

Kessler, R. C., Andrews, G., Mroczek, D., Ustun, B., and Wittchen, H.-U. (1998). The World Health Organization composite international diagnostic interview short-form (CIDI-SF). *International journal of methods in psychiatric research*, 7(4):171–185.

Kostro, D., Abdulkadir, A., Durr, A., Roos, R., Leavitt, B. R., Johnson, H., Cash, D., Tabrizi, S. J., Scahill, R. I., Ronneberger, O., et al. (2014). Correction of inter-scanner and within-subject variance in structural mri based automated diagnosing. *NeuroImage*, 98:405–415.

Kvålseth, T. O. (2017). On Normalized Mutual Information: Measure Derivations and Properties. *Entropy*, 19(11).

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, 4:863.

Lambiotte, R., Delvenne, J.-C., and Barahona, M. (2014). Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks. *Network Science and Engineering, IEEE Transactions on*, 1(2):76–90.

Lebedeva, A. K., Westman, E., Borza, T., Beyer, M. K., Engedal, K., Aarsland, D., Selbaek, G., and Haberg, A. K. (2017). MRI-based classification models in prediction of mild cognitive impairment and dementia in late-life depression. *Frontiers in aging neuroscience*, 9:13.

Lezak, M. (1995). Neuropsychological Assessment, 3rd Edn New York: Oxford University Press.

Liu, Z. and Barahona, M. (2020). Graph-based data clustering via multiscale community detection. *Applied Network Science*, 5(1):1–20.

Malhi, G. and Mann, J. (2018). Depression. *The Lancet*, 392.

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., and Beckmann, C. F. (2016). Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(5):433–447

Merz, E. C., He, X., Noble, K. G., et al. (2018). Anxiety, depression, impulsivity, and brain structure in children and adolescents. *NeuroImage: Clinical*, 20:243–251.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536.

More, S., Eickhoff, S. B., Caspers, J., and Patil, K. R. (2021). Confound removal and normalization in practice: A neuroimaging based sex prediction case study. *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, 12461:3.

Mwangi, B., Ebmeier, K. P., Matthews, K., and Douglas Steele, J. (2012). Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain*, 135(5):1508–1521.

Navrady, L., Wolters, M., MacIntyre, D., Clarke, T., Campbell, A., Murray, A., Evans, K., Seckl, J., Haley, C., Milburn, K., et al. (2018). Cohort profile: STratifying Resilience And Depression Longitudinally (STRADL): a questionnaire follow-up of Generation Scotland: Scottish Family Health Study (GS: SFHS). *International journal of epidemiology*, 47(1):13–14g.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577.

Niida, R., Yamagata, B., Matsuda, H., Niida, A., Uechi, A., Kito, S., and Mimura, M. (2019). Regional brain volume reductions in major depressive disorder and bipolar disorder: An analysis by voxel-based morphometry. *International journal of geriatric psychiatry*, 34(1):186–192.

Nugent, A. C., Davis, R. M., Zarate Jr, C. A., and Drevets, W. C. (2013). Reduced thalamic volumes in major depressive disorder. *Psychiatry Research: Neuroimaging*, 213(3):179–185.

Panizzon, M. S., Fennema-Notestine, C., Eyler, L. T., Jernigan, T. L., Prom-Wormley, E., Neale, M., Jacobson, K., Lyons, M. J.,Grant, M. D., Franz, C. E., et al. (2009). Distinct genetic influences on cortical surface area and cortical thickness. *Cerebral cortex*, 19(11):2728–2735.

Patel, M. J., Andreescu, C., Price, J. C., Edelman, K. L., Reynolds III, C. F., and Aizenstein, H. J. (2015). Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *International journal of geriatric psychiatry*, 30(10):1056–1067.

Qi, Q., Wang, W., Deng, Z., Weng, W., Feng, S., Li, D., Wu, Z., and Liu, H. (2018). Gray matter volume abnormalities in the reward system in first-episode patients with major depressive disorder. In *International Conference on Advanced Machine Learning Technologies and Applications*, pages 704–714. Springer.

Qiao, J., Tao, S., Wang, X., Shi, J., Chen, Y., Tian, S., Yao, Z., and Lu, Q. (2020). Brain functional abnormalities in the amygdala subregions is associated with anxious depression. *Journal of Affective Disorders*, 276:653–659.

Qiu, L., Huang, X., Zhang, J., Wang, Y., Kuang, W., Li, J., Wang, X., Wang, L., Yang, X., Lui, S., et al. (2014). Characterization of major depressive disorder using a multiparametric classification approach based on high resolution structural images. *Journal of psychiatry & neuroscience: JPN*, 39(2):78.

Ran, S., Zuo, Z., Li, C., Yin, X., Qu, W., Tang, Q., Wang, Y., Shi, Y., and Li, H. (2020). Atrophic Corpus Callosum Associated with Altered Functional Asymmetry in Major Depressive Disorder. *Neuropsychiatric Disease and Treatment*, 16:1473.

Raven, J., Raven, J., Raven, J. C., et al. (2003). Manual for Raven's progressive matrices and vocabulary scales: Section 1: general overview.

Ritchie, K., Allard, M., Huppert, F. A., Nargeot, C., Pinek, B., and Ledesert, B. (1993). Computerized cognitive examination of the elderly (ECO): the development of a neuropsychological examination for clinic and population use. *International Journal of Geriatric Psychiatry*, 8(11):899–914.

Romaniuk, L., Sandu, A.-L., Waiter, G. D., McNeil, C. J., Xueyi, S., Harris, M. A., Macfarlane, J. A., Lawrie, S. M., Deary, I. J., Murray, A. D., et al. (2019). The neurobiology of personal control during reward learning and its relationship to mood. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(2):190–199.

Rupprechter, S., Romaniuk, L., Series, P., Hirose, Y., Hawkins, E., Sandu, A.-L., Waiter, G. D., McNeil, C. J., Shen, X., Harris, M. A., Campbell, A., Porteous, D., Macfarlane, J. A., Lawrie, S. M., Murray, A. D., Delgado, M. R., McIntosh, A. M., Whalley, H. C., and Steele, J. D. (2020). Blunted medial prefrontal cortico-limbic reward-related effective connectivity and depression. *Brain*, 143(6):1946–1956.

Schaub, M. T., Delvenne, J.-C., Yaliraki, S. N., and Barahona, M. (2012a). Markov Dynamics as a Zooming Lens for Multiscale Community Detection: Non Clique-Like Communities and the Field-of-View Limit. *PLOS ONE*, 7(2):1–11.

Schaub, M. T., Delvenne, J.-C., Yaliraki, S. N., and Barahona, M. (2012b). Markov Dynamics as a Zooming Lens for Multiscale Community Detection: Non Clique-Like Communities and the Field-of-View Limit.

Schmaal, L., Hibar, D., Sämann, P., Hall, G., Baune, B., Jahanshad, N., Cheung, J., Van Erp, T., Bos, D., Ikram, M., et al. (2017). Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Molecular psychiatry*, 22(6):900–909.

Schmaal, L., Veltman, D. J., van Erp, T. G., Sämann, P., Frodl, T., Jahanshad, N., Loehrer, E., Tiemeier, H., Hofman, A., Niessen, W., et al. (2016). Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Molecular psychiatry*, 21(6):806–812.

Shen, X., Reus, L. M., Cox, S. R., Adams, M. J., Liewald, D. C., Bastin, M. E., Smith, D. J., Deary, I. J., Whalley, H. C., and McIntosh, A. M. (2017). Subcortical volume and white matter integrity abnormalities in major depressive disorder: findings from UK Biobank imaging data. *Scientific Reports*, 7(1):1–10.

Smith, B. H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., Kerr, S. M., Deary, I. J., MacIntyre, D. J., Campbell, H., McGilchrist, M., et al. (2012). Cohort Profile: Generation Scotland: Scottish Family Health Study (GS: SFHS). The study, its participants and their potential for genetic research on health and illness. *International journal of epidemiology*, 42(3):689–700.

Snoek, L., Miletić, S., and Scholte, H. S. (2019). How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*, 184:741–760.

Stolicyn, A., Harris, M. A., Shen, X., Barbu, M. C., Adams, M. J., Hawkins, E. L., de Nooij, L., Yeung, H. W., Murray, A. D., Lawrie, S. M., et al. (2020). Automated classification of depression from structural brain measures across two independent community-based cohorts. *Human brain mapping*, 41(14):3922–3937.

Suh, J. S., Schneider, M. A., Minuzzi, L., MacQueen, G. M., Strother, S. C., Kennedy, S. H., and Frey, B. N. (2019). Cortical thickness in major depressive disorder: a systematic review and meta-analysis. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 88:287–302.

Sullivan, G. M. and Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of graduate medical education*, 4(3):279.

Tokuda, T., Yoshimoto, J., Shimizu, Y., Okada, G., Takamura, M., Okamoto, Y., Yamawaki, S., and Doya, K. (2018). Identification of depression subtypes and relevant brain regions using a data-driven approach. *Scientific reports*, 8(1):1–13.

Treadway, M. T., Waskom, M. L., Dillon, D. G., Holmes, A. J., Park, M. T. M., Chakravarty, M. M., Dutra, S. J., Polli, F. E., Iosifescu, D. V., Fava, M., et al. (2015). llness progression, recent stress, and morphometry of hippocampal subfields and medial prefrontal cortex in major depression. *Biological psychiatry*, 77(3):285–294.

van Velzen, L. S., Kelly, S., Isaev, D., Aleman, A., Aftanas, L. I., Bauer, J., Baune, B. T., Brak, I. V., Carballedo, A., Connolly, C. G., et al. (2020). White matter disturbances in major depressive disorder: a coordinated analysis across 20 international cohorts in the enigma mdd working group. *Molecular psychiatry*, 25(7):1511–1525.

Warne, R. T. and Burningham, C. (2019). Spearman's g found in 31 non-western nations: Strong evidence that g is a universal phenomenon. *Psychological Bulletin*, 145(3):237.

Webb, C. A., Weber, M., Mundy, E. A., and Killgore, W. D. (2014). Reduced gray matter volume in the anterior cingulate, orbitofrontal cortex and thalamus as a function of mild depressive symptoms: a voxel-based morphometric analysis. *Psychological medicine*, 44(13):2833.

Wechsler, D. Psychological Corporation; London, UK: 1998. *WAIS-III UK Wechsler Adult Intelligence Scale. [Google Scholar]*.

Wechsler, D., Wycherly, R., Benjamin, L., Crawford, J., and Mockler, D. (1998). Wechsler Memory Scale-III. *London: The Psychological Corporation Limited*.

Wei, D., Wang, K., Meng, J., Zhuang, K., Chen, Q., Yan, W., Xie, P., and Qiu, J. (2020). The reductions in the subcallosal region cortical volume and surface area in major depressive disorder across the adult life span. *Psychological medicine*, 50(3):422–430.

Weir, W. H., Emmons, S., Gibson, R., Taylor, D., and Mucha, P. J. (2017). Post-processing partitions to identify domains of modularity optimization. *Algorithms*, 10(3):93.

Winkler, A. M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P. T., Duggirala, R., and Glahn, D. C. (2010). Cortical thicknessor grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *Neuroimage*, 53(3):1135–1146.

Ye, J., Wu, C., Chu, X., Wen, Y., Li, P., Cheng, B., Cheng, S., Liu, L., Zhang, L., Ma, M., et al. (2020). Evaluating the effect of birth weight on brain volumes and depression: An observational and genetic study using UK Biobank cohort. *European Psychiatry*, 63(1).

Zhao, K., Liu, H., Yan, R., Hua, L., Chen, Y., Shi, J., Lu, Q., and Yao, Z. (2017). Cortical thickness and subcortical structure volume abnormalities in patients with major depression with and without anxious symptoms. *Brain and behavior*, 7(8):e00754.

Zhou, Y., Zhao, L., Zhou, N., Zhao, Y., Marino, S., Wang, T., Sun, H., Toga, A. W., and Dinov, I. D. (2019). Predictive Big Data Analytics using the UK Biobank Data. *Scientific reports*, 9(1):6012.

Table 1 Participants Characteristics for the two studied cohorts.

Table 1a Participants Characteristics for the GS subsample dataset

GS subsample	Mean(SD)	N
Age	59.94(9.787)	980
Sex(F:M)	602:378	980
Site(Aberdeen:Dundee)	516:464	980
ICV* (in cm^3)	1397 (224.2)	980
MDD(Control:Cases)	677:302	979
DSy	68.86(14.98)	879
VF	43.05(11.89)	879
MHV	31.69(4.065)	879
Matrix	8.325(2.411)	879
LM	31.67(7.250)	873

Note: The ICV here was not standardised for each site. DSy: Digit-symbol Coding, VF: Verbal Fluency Total Score, MHV: Mill Hill Vocabulary, Matrix: Matrix Reasoning, LM: Logical Memory.

Table 1b Participants Characteristics for the UKB dataset

UKB	Mean(SD)	N
Age	62.47(7.464)	8900
Sex(F:M)	4682:4218	8900
Site	Manchester	8900
ICV (in cm^3)	1519(147.8)	8900
MDD(Control:Cases)	3865:1658	5523
VNR	6.8811(2.097)	8484
RT (in log x)	6.3582(0.1687)	8796
Pairs Match (in log (x+1))	1.288(0.6473)	8836
Prospective Memory(1:0)	7885:936	8821

Note: VNR: Verbal Numerical Reasoning. RT: Reaction Time. Pairs Match: Pairs Matching. For the prospective memory test, 1 means recall at the first attempt and 0 otherwise.

Note: N is the number of participants for whom data is available. Age is in years. Measures for DSy/ VF / MHV / Matrix / LM and VNR represent raw task scores. The x in RT and Pairs Match represent raw task scores.

Table 2 NMI between 2-cluster partitions of the different k-NN graphs and the 2-cluster partitions of the 11-NN graph in GS subsample for each of cortical thickness (CT), cortical surface area (CSA), cortical volume (CV) and subcortical volume (subCV) measures.

Similarity Measure	5-NN Graph	7-NN Graph	9-NN Graph
CT NMI (Accuracy)	0.7405(95.5%)	0.8014(96.6%)	0.7796(95.9%)
CSA NMI (Accuracy)	0.5618(89.7%)	0.5951(89.6%)	0.7776(96.1%)
CV NMI (Accuracy)	0.7375(95.5%)	0.8284(97.4%)	0.8581(97.9%)
subCV NMI (Accuracy)	0.6664(94.9%)	0.7558(95.9%)	0.7194(94.7%)

Note: The high NMI and accuracies (except for CSA where small fluctuations were seen) indicate that the clustering results of different graphs are similar and hence that the partitioning result is meaningful.

Table 3 NMI between 2-cluster partitions of the different k-NN graphs and the 2-cluster partitions of the 5-NN graph in UKB for each of cortical thickness (CT), cortical surface area (CSA), cortical volume (CV) and subcortical volume (subCV) measures.

Similarity Measure	7-NN Graph	9-NN Graph	11-NN Graph
CT NMI(Accuracy)	0.7883(96.7%)	0.7760(96.6%)	0.7325(95.5%)
CSA NMI(Accuracy)	0.7592(96.0%)	0.7128(95.9%)	0.7038(95.8%)
CV NMI(Accuracy)	0.7623(96.3%)	0.7163(95.2%)	0.6441(92.7%)
subCV NMI(Accuracy)	0.6221(93.4%)	0.6310(92.9%)	0.6290(92.9%)

Note: The high NMI and accuracies indicate that the clustering results of different graphs are similar and hence that the partitioning result is meaningful.

Table 4 Top 20 regions in CT, CSA and CV, and top 10 regions in subCV driving the separation between the clusters for each of the four modalities in both cohorts.

Top Regions in each cohort	GS subsample	Cohen's *d*	UKB	Cohen's *d*
Cortical Thickness	**rh.supramarginal**	**1.6617**	**rh.inferiorparietal**	**1.5359**
	rh.inferiorparietal	**1.6615**	**lh.supramarginal**	**1.4962**
	lh.superiorfrontal	**1.6055**	**rh.precuneus**	**1.4790**
	lh.inferiorparietal	**1.5852**	**rh.supramarginal**	**1.4502**
	rh.precentral	**1.5828**	**lh.superiorfrontal**	**1.4499**
	rh.superiorfrontal	**1.5617**	**lh.inferiorparietal**	**1.4262**
	lh.supramarginal	**1.5517**	**lh.precuneus**	**1.3948**
	rh.superiortemporal	**1.4962**	**lh.rostralmiddlefrontal**	**1.3930**
	lh.precuneus	**1.4942**	**rh.superiorparietal**	**1.3923**
	rh.precuneus	**1.4922**	**rh.superiorfrontal**	**1.3916**
	lh.caudalmiddlefrontal	**1.4836**	**lh.precentral**	**1.3665**
	rh.superiorparietal	**1.4666**	**lh.superiorparietal**	**1.3369**
	lh.precentral	**1.4665**	**lh.caudalmiddlefrontal**	**1.3297**
	rh.caudalmiddlefrontal	**1.4011**	**rh.precentral**	**1.2939**
	rh.postcentral	1.3938	**lh.postcentral**	**1.2761**
	lh.superiorparietal	**1.3852**	**rh.superiortemporal**	**1.2668**
	lh.postcentral	**1.3813**	**rh.rostralmiddlefrontal**	**1.2667**
	lh.superiortemporal	**1.3561**	**lh.superiortemporal**	**1.2567**
	rh.rostralmiddlefrontal	**1.3506**	**rh.caudalmiddlefrontal**	**1.2554**
	lh.rostralmiddlefrontal	**1.3395**	rh.middletemporal	1.2468
Cortical Surface Area	**lh.rostralmiddlefrontal**	**1.3871**	**lh.superiorfrontal**	**1.0902**
	rh.superiorfrontal	**1.3132**	**rh.superiorfrontal**	**1.0301**
	rh.rostralmiddlefrontal	**1.3077**	**lh.rostralmiddlefrontal**	**1.0281**
	rh.middletemporal	**1.2980**	**rh.rostralmiddlefrontal**	**1.0034**
	lh.superiorfrontal	**1.2216**	**rh.superiortemporal**	**0.9897**
	lh.lateralorbitofrontal	**1.2098**	**lh.lateralorbitofrontal**	**0.9896**
	lh.superiortemporal	**1.2075**	**lh.precuneus**	**0.9742**
	rh.superiortemporal	**1.2065**	**lh.superiortemporal**	**0.9596**
	rh.inferiortemporal	**1.1435**	rh.lateralorbitofrontal	0.9054
	rh.precuneus	**1.1261**	**rh.middletemporal**	**0.9047**
	rh.inferiorparietal	1.1009	**rh.precuneus**	**0.8879**
	lh.precuneus	**1.0989**	rh.fusiform	0.8317
	lh.middletemporal	**1.0980**	**lh.postcentral**	**0.8288**
	lh.fusiform	**1.0861**	rh.medialorbitofrontal	0.8186
	lh.inferiortemporal	1.0761	**lh.fusiform**	**0.8068**
	lh.postcentral	**1.0752**	**rh.inferiortemporal**	**0.8035**
	lh.lateraloccipital	1.0652	**lh.middletemporal**	**0.8031**
	lh.supramarginal	1.0649	rh.postcentral	0.7926
	rh.superiorparietal	1.0327	lh.insula	0.7890
	lh.precentral	1.0322	lh.superiorparietal	0.7771
Cortical Volume	**lh.superiorfrontal**	**1.4611**	**lh.precuneus**	**1.0578**
	rh.superiorfrontal	**1.4606**	**rh.superiorfrontal**	**1.0278**
	lh.lateralorbitofrontal	**1.4279**	**lh.superiorfrontal**	**1.0213**
	rh.precuneus	**1.3651**	**rh.precuneus**	**1.0201**
	rh.superiortemporal	**1.2925**	**lh.lateralorbitofrontal**	**0.9902**
	lh.precuneus	**1.2347**	**rh.superiortemporal**	**0.9554**
	lh.superiortemporal	**1.1974**	**lh.rostralmiddlefrontal**	**0.9449**
	lh.rostralmiddlefrontal	**1.1870**	**lh.precentral**	**0.9142**
	lh.supramarginal	**1.1595**	**lh.superiortemporal**	**0.9087**
	rh.insula	1.1360	**rh.precentral**	**0.9072**
	lh.insula	**1.1316**	**rh.rostralmiddlefrontal**	**0.9070**
	rh.rostralmiddlefrontal	**1.1303**	**rh.lateralorbitofrontal**	**0.8954**
	lh.lateraloccipital	1.1243	**lh.insula**	**0.8720**
	rh.lateralorbitofrontal	**1.1239**	**rh.inferiorparietal**	**0.8444**
	rh.inferiorparietal	**1.0981**	lh.postcentral	0.8406
	rh.superiorparietal	**1.0974**	**rh.superiorparietal**	**0.8341**

	lh.precentral	**1.0793**	**lh.supramarginal**	**0.8265**
	rh.precentral	**1.0691**	**lh.superiorparietal**	**0.8171**
	lh.superiorparietal	**1.0581**	rh.lateraloccipital	0.7796
	rh.middletemporal	1.0376	lh.inferiorparietal	0.7776
	Right.VentralDC	**1.7621**	**Right.VentralDC**	**1.4164**
	Left.VentralDC	**1.6753**	**Left.VentralDC**	**1.2586**
	Brain.Stem	**1.6241**	**Brain.Stem**	**1.2144**
	Right.Thalamus.Proper	**1.4319**	**Left.Cerebellum.White.Matter**	**1.0156**
Subcortical Volume	Right.Cerebellum.Cortex	1.2596	**Right.Cerebellum.White.Matter**	**1.0027**
	Left.Cerebellum.Cortex	1.2545	Right.Putamen	0.9641
	Left.Thalamus.Proper	1.2382	Left.Putamen	0.8870
	Right.Hippocampus	**1.1968**	Right.Pallidum	0.8345
	Left.Cerebellum.White.Matter	**1.1537**	**Right.Hippocampus**	**0.8142**
	Right.Cerebellum.White.Matter	**1.1153**	**Right.Thalamus.Proper**	**0.7929**

Note: The overlapping regions across cohorts are in bold text. lh: left hemisphere, rh: right hemisphere.

Table 5 p-values and Cohen's d values (in brackets) for cluster effects on each cognitive performance measure for the four measure modalities in GS subsample.

Structural brain measure modality	DSy	VF	MHV	Matrix	Memory	g-factor
CTs	1.68e-5*(**0.3455**)	0.0914(-0.1431)	0.3291(-0.0591)	0.0051*(**0.1971**)	0.2418(0.1337)	0.2220(0.1154)
CSAs	4.26e-5*(**0.3181**)	0.6627(0.0717)	0.1172(0.1092)	1.56e-5*(**0.3229**)	0.2092(0.0965)	8.69e-5*(**0.2771**)
CVs	2.06e-8*(**0.4161**)	0.9541(0.0239)	0.4300(0.0909)	1.81e-10*(**0.4450**)	0.0106*(**0.2075**)	5.27e-6*(**0.3490**)
subCVs	4.19e-7*(**0.4036**)	0.3231(0.0867)	0.7565(0.0022)	3.88e-4*(**0.3057**)	0.3155(0.1085)	0.0022*(**0.2529**)

Note: DSy=WAIS-III Digit Symbol Coding, VF = Verbal Fluency, MHV = Mill Hill Vobaulary, Matrix=Matrix Reasoning total correct, g-factor = general intelligence coefficient derived from cognitive variables. p-values were FDR-corrected. Significant positive effect sizes in FreeSurfer measures were related to positive effect sizes in cognitive measures.

Table 6 p-values and Cohen's d values (in brackets) for cluster effects on each cognitive performance measure for the four measure modalities in UKB.

Structural brain measure modality	VNR	RT	Pairs Match	ProsMemory	*g*-factor
CTs	0.0318*(**0.0502**)	0.5595(-0.0230)	0.3748(-0.0258)	0.0489*(**0.0485**)	0.0318*(**0.0573**)
CSAs	5.19e-5*(**0.0999**)	0.7270(0.0049)	0.1637(-0.0298)	0.5052(0.0178)	2.05e-4*(**0.0868**)
CVs	1.17e-5*(**0.1091**)	0.8778(-0.0055)	0.1324(-0.0308)	0.0468*(**0.0483**)	1.24e-5*(**0.1070**)
subCVs	4.48e-5*(**0.0923**)	0.0011*(**-0.0716**)	0.5062(0.0121)	0.5062(0.0156)	4.48e-5*(**0.0919**)

Note: VNR = Verbal Numerical Reasoning. RT = Reaction Time. Pairs Match = Pairs Matching. ProsMemory = Prospective Memory. g-factor = general intelligence coefficient derived from all cognitive variables. p-values were FDR-corrected. Similar to GS subsample, significant positive effect sizes in UKB FreeSurfer measures were related to positive effect sizes in cognitive measures, except for reaction time task.
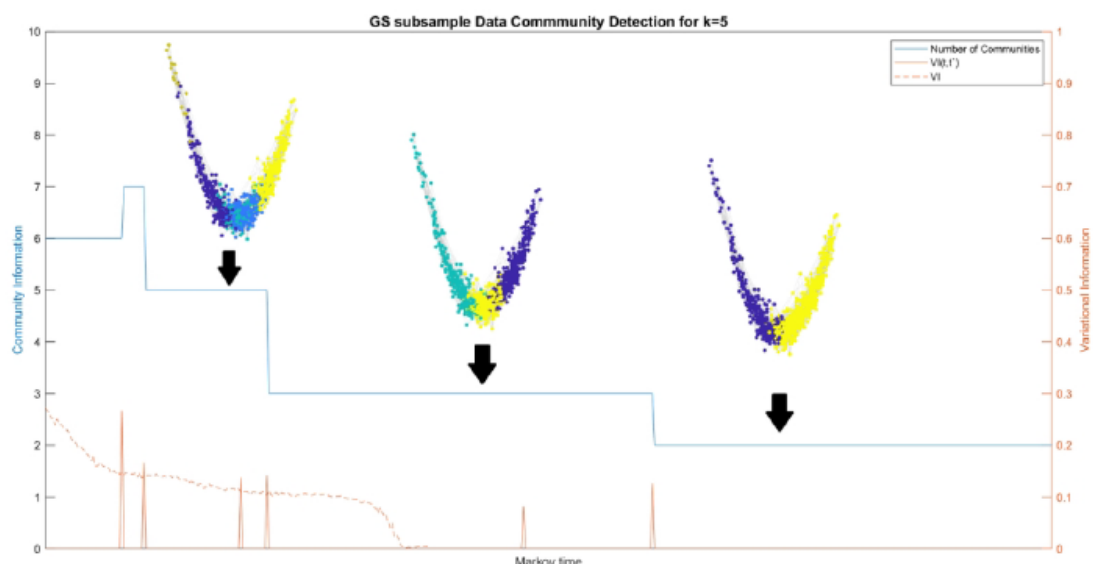
Figure 1: Community information, VI(t,t'), together with 2D projections of the graphs, with nodes coloured according to cluster labels and nodes arranged with reference to 5-cluster partition, over Markov time for the clustering analysis on 5-NN graph, constructed with residualised FreeSurfer surface area metrics for GS subsample. Clustering results are based on time-dependent Markov Stability optimisation and had undergone post-processing for smoothing stability and obtaining more stable clustering result. The fact that the VI(t,t') stays zero for most of the timesteps indicates robustness of the partition results. Note that the 500 time-steps are logarithmically spaced from 1 to 100
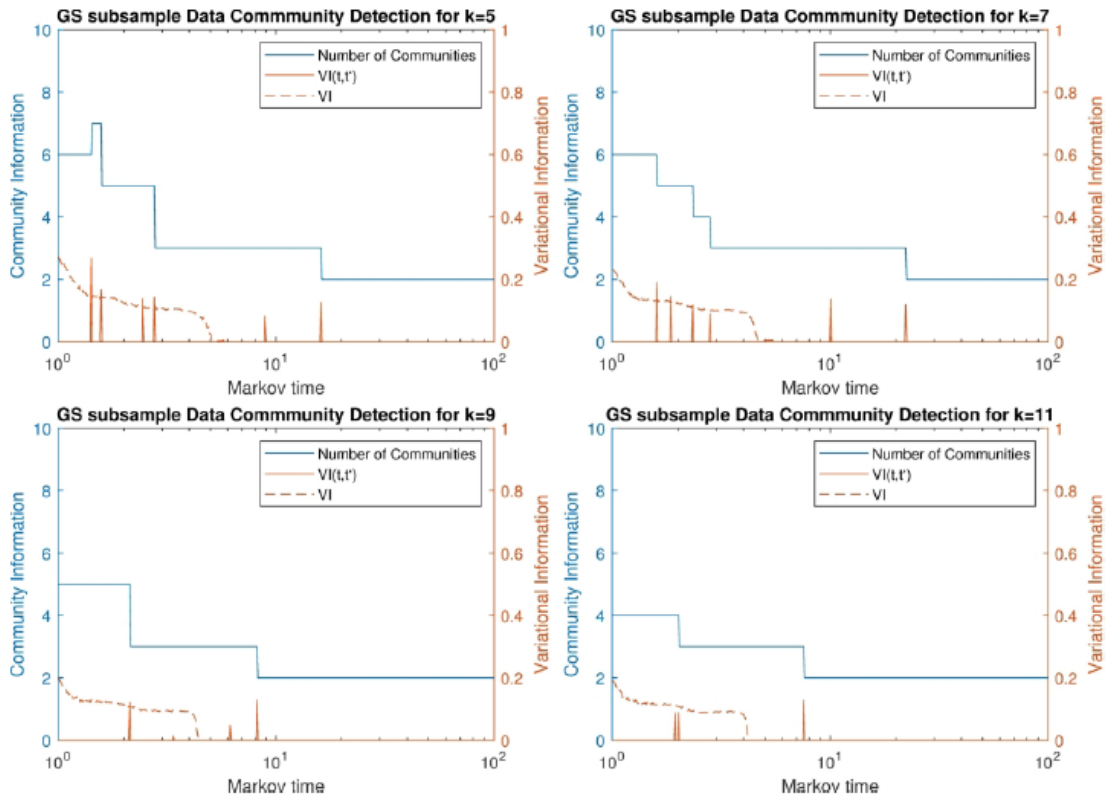
Figure 2: Number of clusters found (left y-axis) and VI(t,t') (right y-axis) over Markov time for k-NN graphs based on the FreeSurfer surface area metrics in GS subsample. Low VI indicates that variations among different Louvain runs are small, which implies that the important clusters are from partitions into 2 or 3 modules. The VI(t,t') is an evaluation of the variational information of the partition at timestep tn with that of tn−1 and the spikes indicate changes in partitioning result. No spikes are present in any of the plots after clustering into 2, implying the 2 clusters are equally stable with any k. Since the plateau at 2-cluster partition was the largest (i.e. the algorithm stays at 2-cluster partition for the largest time period), we therefore identified that the 2-cluster partition as the most stable partition.
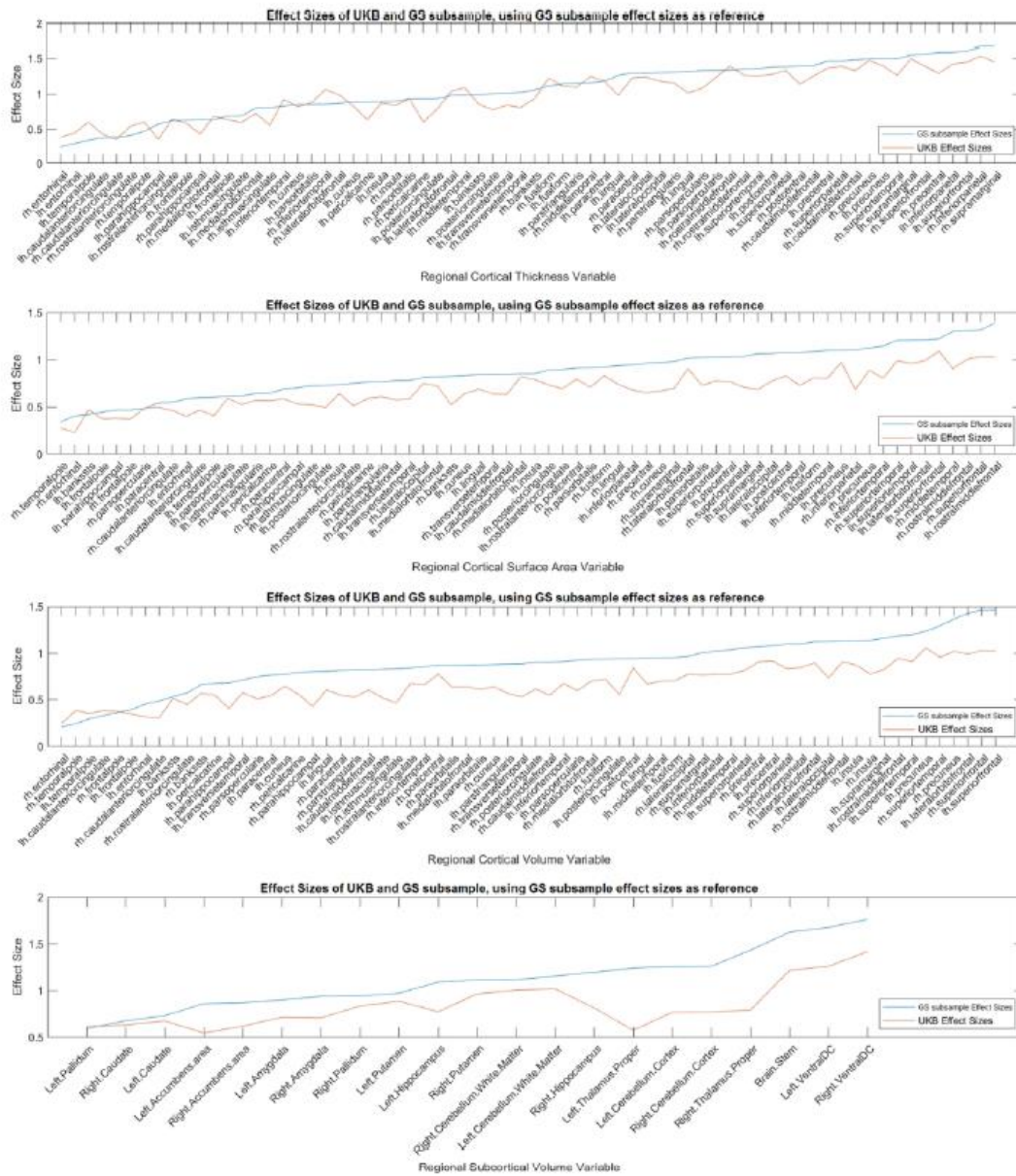
Figure 3: Mean differences between the two clusters in CT, CSA, CV and subCV measures for GS subsample and UKB, with ordered GS subsample regional Cohen's d values(blue lines) as reference, where x-axes are the FreeSurfer regions (there are 68 cortical regions and 21 subcortical regions). High correlation indicates successful replication and that the clustering algorithm has likely identified natural groupings for general population.
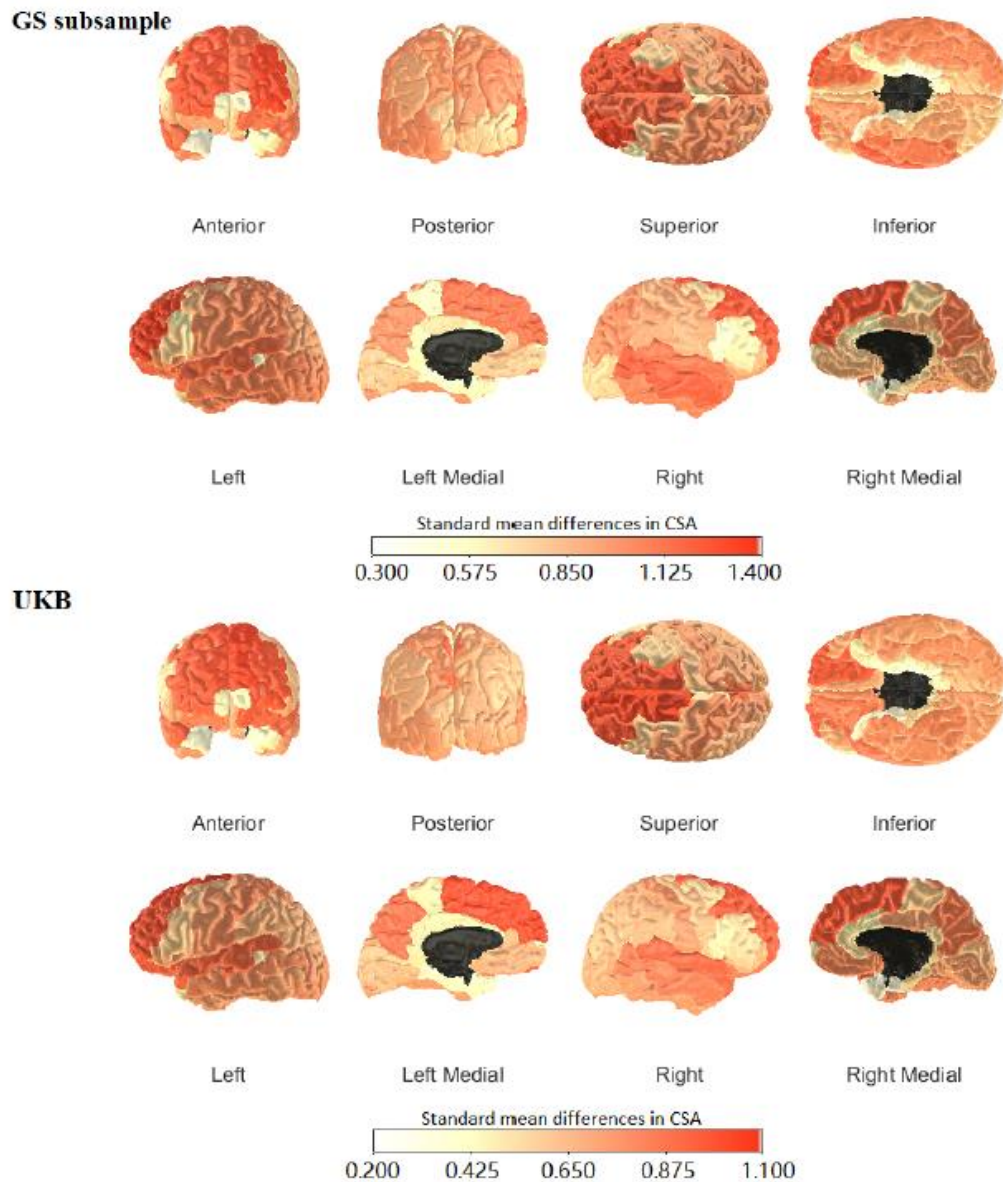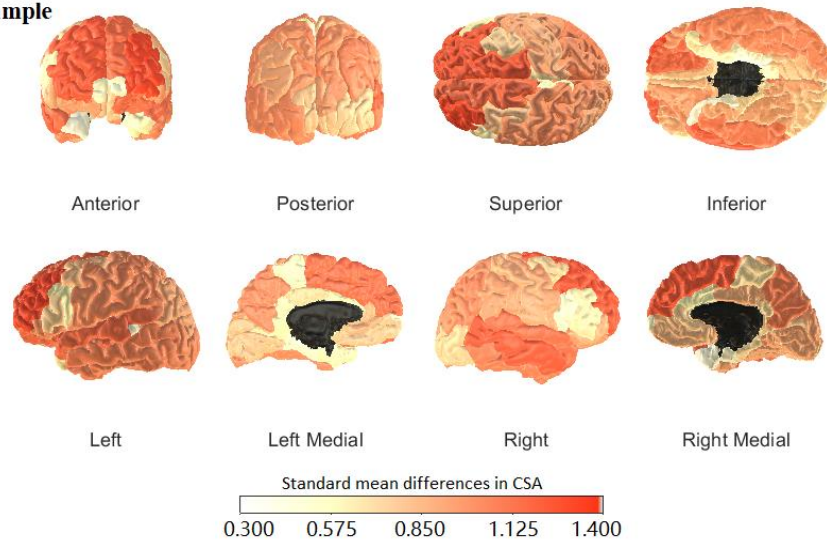
Figure 4: Standardised mean differences in regional CSAs between the two clusters identified in the GS subsample and in UKB. Higher differences between clusters were found for lateral orbitofrontal, post central, precentral, precuneus, rostral middle frontal, superior frontal, superior parietal and supramarginal areas in both hemispheres. Most of these areas are closely related to executive functions and decision making. These regions are also associated with various diseases that may contribute to larger differences between individuals.
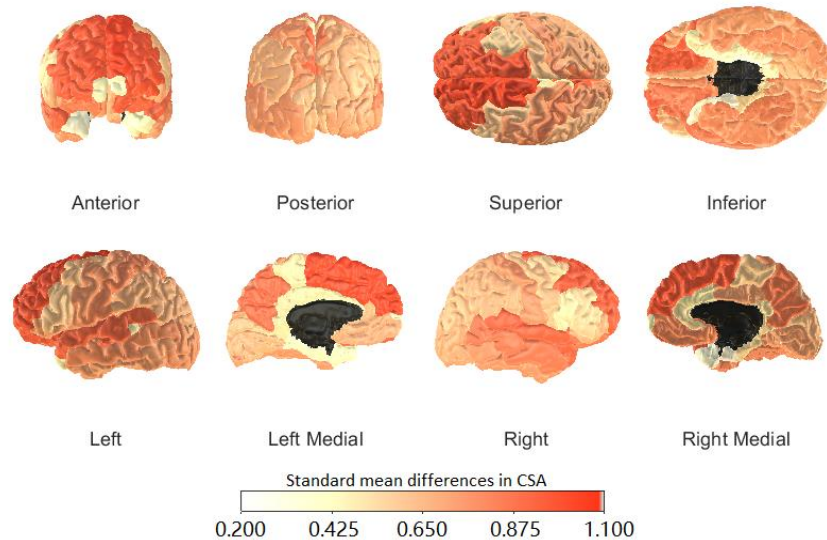
Spectral Clustering Based on Structural Magnetic Resonance Imaging and its Relationship with Major Depressive Disorder and Cognitive Ability

Hon Wah Yeung [1], Xueyi Shen[1], Aleks Stolicyn[1], Laura De Nooij[1], Mathew A. Harris[1], Liana Romaniuk[1], Colin R. Buchanan[2], Gordon D. Waiter[3], Anca-Larisa Sandu[3], Christopher J. McNeil[3], Alison Murray[3], J. Douglas Steele[4], Archie Campbell[5], David Porteous[5,6], Stephen M. Lawrie[1], Andrew M. McIntosh[1,5], Simon R. Cox[2], Keith M. Smith[7,8], and Heather C. Whalley[1]

We thoroughly study the existence and replicability of natural groupings of subjects from structural MRI measurements and their relationship to cognition and Major Depressive Disorder (MDD). To do this, we employ an advanced clustering algorithm to nearest neighbour networks of structural similarities, corrected for age and gender. We find replicable clusters across two large datasets which are associated with cognition but find no associations with MDD.