Cognitive Subtypes of Psychiatric Disorders and Associations with Brain Structure in the UK Biobank: A Machine Learning Approach

Joshua Unrau

ID: 260856488

Department of Psychology, McGill University

PSYC 498: U3 Honours Research Project

Principal Investigator: Dr. Martin Lepage

Co-Supervisor: Dr. Katie Lavigne

McGill University

12 April 2022

Abstract

Neurocognitive deficits (e.g., impairments in attention, memory, or executive function) are core features of mood and psychotic disorders. Identifying putatively more homogeneous cognitive clusters in a transdiagnostic sample could serve as a preliminary step towards developing a taxonomy that more accurately reflects the pathophysiology underlying these impairments. Here, we used the k-means algorithm to derive cognitive clusters in a sample from the UK Biobank consisting of 680 individuals with mood or psychotic disorders and 680 age and sex matched controls (N=1360). Subsequently, we evaluated the performance of a variety of supervised learning algorithms in predicting derived clusters using measures of cortical morphometry (i.e., volume, thickness, and surface area). Models were trained on 75% of cases using stratified fivefold cross validation and hyperparameter tuning was performed using Bayesian optimization. Two cognitive clusters were identified: high performing (n=697, 51.3%) and low performing (n=663, 48.7%). The high-performing cluster outperformed the low-performing cluster on all cognitive tasks. Using a random forest classifier, we were able to predict clusters with an overall accuracy of 80% on training data (AUC = .90) and 79% on holdout data (AUC = .89). The most informative features included right middle temporal gyrus area, right caudal middle frontal gyrus thickness, left inferior temporal gyrus volume, left insula thickness, and right superior frontal gyrus thickness. These findings demonstrate that cortical morphometry can reliably distinguish cognitive subtypes in a mixed sample including significant numbers of patients with affective disorders and healthy controls.

Subject Area: Clinical Psychology

Keywords: Brain, Cognition, Clusters, Predict, Subtypes

Cognitive Subtypes of Psychiatric Disorders and Associations with Brain Structure in the UK Biobank: A Machine Learning Approach

Neurocognitive deficits (e.g., impairments in attention, memory, or executive function) have long been recognized to be core features of schizophrenia. Indeed, both Bleuler (1911) and Kraepelin (1919) emphasized the centrality of cognitive impairments in dementia praecox (for a review, see Green & Harvey, 2014). It is now well-established that patients with schizophrenia spectrum disorders (SSDs) exhibit widespread deficits across a range of cognitive domains (Millan et al., 2012). In a meta-analysis of 204 studies, Heinrichs and Zakzanis (1998) assessed the magnitude of cognitive impairment in SSD patients on 22 cognitive tests, concluding that SSD patients exhibit moderate to large impairments on all metrics compared to healthy controls (Cohen's d range: -0.62 to -1.53).

While cognitive impairments have also been noted in affective disorders since the time of Kraepelin (1921), such impairments have historically been viewed as a reversible and largely secondary phenomenon (McAllister, 1981). In recent years, however, several lines of evidence have begun to converge casting serious doubts on this view. First, numerous studies have called into question the notion that cognitive deficits are an effective means of distinguishing between affective and non-affective psychoses (Bora et al., 2009; Hill et al., 2013; Vöhringer et al., 2013). Indeed, a growing body of evidence indicates that cognitive deficits in affective psychoses are qualitatively similar to those in SSDs, suggesting a continuum of impairment, rather than a dichotomy (Bora et al., 2009; Hill et al., 2013). Second, there is also reason to believe that this continuum extends beyond psychotic disorders. Although psychotic symptoms in bipolar disorder (BD) are associated with reduced global cognitive performance (Cohen's d: -0.22), the presence of these symptoms does not appear to be associated with a categorically different cognitive profile (Bora et al., 2010). These findings are supported by a systematic review concluding that euthymic BD patients report a similar profile of cognitive impairments to SSD patients, albeit of a generally milder severity (Vöhringer et al., 2013). Indeed, Krabbendam et al. (2005) reported that remitted SSD patients exhibited significantly poorer cognitive performance across nine of eleven cognitive domains compared to remitted BD patients (Cohen's d range: -0.34 to -0.63). Third, although historically the cognitive impairments associated with depression received little to no attention (Perini et al., 2019), there is now widespread acknowledgement that cognitive impairment constitutes a core feature of depressive pathology (Pan et al., 2019; Perini et al., 2019; Rock et al., 2014). In a meta-analysis, Rock et al. (2014) observed that current MDD patients exhibited significantly poorer cognitive performance compared to healthy controls with respect to executive function (Cohen's d range: −0.34 to −0.54), memory (Cohen's d range: −0.41 to −0.50) and attention (Cohen's d = −0.65). Importantly, compared to controls, patients with remitted depressive symptoms continued to experience significantly impaired attention and executive function (Cohen's d range: −0.52 to −0.61), and demonstrated a trend towards impaired memory (Cohen's d range: −0.22 to −0.54). Another recent meta-analysis of 252 studies confirmed the presence of cognitive deficits in patients with remitted depressive symptoms. Semkovska et al. (2019) observed that remitted MDD patients exhibited significantly poorer selective attention, working memory, and long-term memory than healthy controls. Given the absence of confounding symptomatology, these findings support the contention that cognitive impairments are intrinsic to depressive pathology, rather than chiefly a secondary symptom. Considering the substantial impact that these impairments have on patient functioning and quality of life (Perini et al., 2019), the development of more effective treatments is paramount.

To this end, one avenue that has received increased attention in recent years is the possibility of developing more targeted interventions using machine learning techniques. Fundamentally, machine learning involves programming computers to learn through experience rather than to follow an explicit set of instructions (Samuel, 1959). Broadly, machine learning models can be categorized as either supervised or unsupervised (Vanderplas, 2016). Supervised learning entails modeling relationships between features (i.e., properties that describe data) and predetermined labels (e.g., patient, control), typically for the purpose of generalizing from known examples (Müller & Guido, 2016; Vanderplas, 2016). Supervised learning models are subdivided into classification tasks, in which the goal is to predict a predefined class label (e.g., predicting whether an individual has been diagnosed with BP or MDD), and regression tasks, which involve predicting a floating-point number (e.g., predicting an individual's performance on an arbitrary cognitive metric). In general, supervised learning models are built using a subset of training data, and their performance is evaluated by their ability to generalize to a holdout set of data that was not used to train the model (for an overview of supervised learning, see Müller & Guido, 2016, Chapter 2). There are a variety of metrics that can be used to assess the performance of a supervised learning model. For example, in classification tasks, accuracy is defined as the number of correct predictions outputted by the model divided by the number of total predictions made (Burkov, 2019, Chapter 5). Unsupervised learning, on the other hand, involves modeling features without known labels (e.g., clustering; Müller & Guido, 2016; Vanderplas, 2016). While unsupervised learning models have proven extremely useful in a variety of applications, the lack of an objective reference point makes evaluating their performance extremely difficult (for an overview of unsupervised learning, see Burkov, 2019, Chapter 9; Müller & Guido, 2016, Chapter 3).

In the context of cognition, unsupervised clustering techniques have garnered particular interest in the literature. Given the considerable heterogeneity in cognitive performance both within and across diagnostic boundaries, delineating relatively more homogeneous cognitive phenotypes may be a viable strategy for improving our understanding of the pathophysiological mechanisms underlying both psychotic and mood disorders (Green et al., 2020). To this end, considerable effort has been devoted to identifying data-driven cognitive subtypes. In a recent meta-analysis, Green et al. (2020) identified 24 studies that attempted to delineate such subtypes in psychotic disorders, 13 of which were in SSD patients, 8 of which were in BD patients and 5 of which were in mixed SSD and BD populations. Interestingly, despite the disparate methodologies across studies, including a range of clustering methods and cognitive domains, nearly all findings converge on several key points; studies have consistently identified a severely impaired subgroup, with moderate to large cognitive impairments compared to healthy controls across all cognitive domains, as well as a cognitively spared subgroup with similar performance to healthy controls (Green et al., 2020; Karantonis et al., 2022).

By comparing neurological markers between these subgroups, we may be able to gain a better understanding of the mechanisms underlying these differences. To this end, several studies have also examined associations between various cognitive subtypes and brain structure in SSD patients (for a systematic review, see Karantonis et al., 2022). Notably, Cobia et al. (2011) found two cognitive subtypes of SSD patients, one of which was characterized by widespread cognitive impairments and reductions in cortical thickness, the other by both near-normal cortical thickness and cognitive performance. Geisler et al. (2015) identified four cognitive subtypes, characterized by (1) poor verbal fluency and processing speed with reduced cortical thickness in Wernicke's area; (2) poor verbal memory and motor control with reduced volume in the right hippocampus;

(3) poor face memory and processing speed with reduced cortical thickness in the lingual gyrus, fusiform face area, occipital face area, superior frontal gyrus, rostral anterior cingulate cortex, and middle temporal gyrus; and (4) poor overall cognitive functioning with generally widespread cortical thinning. In three samples including only SSD patients, only BD patients, and a mixed sample of SSD and BD patients, Van Rheenen et al. (2017) identified three cognitive subtypes: a severely impaired cluster, a mild-moderately impaired cluster and a relatively intact cognitive cluster. In a follow-up study, they examined differences in cortical morphometry between clusters of SSD patients (Van Rheenen et al., 2018). The authors found that when compared to the relatively intact and mild-moderately impaired group, the severely impaired group displayed volumetric reductions in the left lateral orbitofrontal cortex, parahippocampal gyrus, temporal pole, and right pars triangularis, as well as reduced thickness of the left rostral anterior cingulate and the left parahippocampal gyrus (Van Rheenen et al., 2018). Interestingly, Gould et al. (2014) attempted to distinguish two cognitive subtypes from each other using grey and white matter volume with support vector machine classification, a supervised learning technique (for a mathematical overview, see Burkov, 2019, Chapter 3), achieving an accuracy rate of 71%.

Despite the relative abundance of research on psychotic disorders, only a few studies have applied data-driven clustering techniques to non-psychotic mood disorders. Recently, Yang et al. (2021) stratified two subgroups of patients with MDD based on IQ, finding associations between the low IQ group and other impairments in cognitive performance as well as functional network aberrations. Baller et al. (2021) identified three cognitive subtypes in a sample of youth with MDD: high performing (high accuracy, moderate speed), cognitively impaired (low accuracy, slow speed), and impulsive (low accuracy, fast speed). The authors reported significant differences in activation during a working memory task in a variety of regions, including the left anterior dorsolateral prefrontal cortex, dorsal anterior cingulate, left dorsal frontal cortex, as well as the left and right precuneus, which they interpreted as indicating the presence of biologically relevant subtypes. In a mixed sample including healthy controls and patients with both mood and psychotic disorders, Pelin et al. (2021) identified five clusters based on multiple symptom dimensions, including cognition; the authors also used a supervised learning technique to attempt to predict these subtypes using demographic variables and genetic markers.

Ultimately, these findings have established that it is possible to apply supervised learning techniques to predict data-driven clusters. However, we are not aware of any studies that have applied supervised learning to predict cognitive clusters using measures of brain structure in a sample with substantial numbers of patients with non-psychotic mood disorders. Such research could yield new insights into the overlap between psychotic and mood disorders, as well as the overlap between clinical and non-clinical populations. Indeed, including healthy controls in clustering analyses enables assessing the spectrum between well-being and disease (Pelin et al., 2021).

The present study examined the relationship between cognitive subtypes and brain structure in the UK Biobank, an ongoing large prospective cohort study involving over 500,000 individuals (Allen et al., 2012). Our primary aims were: (1) to identify cognitive subtypes using an unsupervised clustering algorithm in a mixed sample of healthy controls and patients with psychotic and mood disorders, and (2) to predict these derived subtypes using measures of cortical morphometry (i.e., volume, thickness, and surface area). We hypothesized that we would identify a three cluster solution as best fitting the data, consistent with findings of Baller et al. (2021) and most studies in psychotic BD patients (Green et al., 2020). In predicting these clusters, we anticipated achieving an accuracy rate above chance, but less than that reported by

Gould et al. (2014). Given that cognitive impairments are well-known to be less severe in affective disorders (East-Richard et al., 2020), we reasoned that the potential mechanisms underlying these differences would be less well distinguished than in psychotic disorders.

Methods

UK Biobank

The UK Biobank is an ongoing large prospective cohort study. Between 2006 and 2010, approximately 9.2 million individuals aged 40–69 who resided within a 25-mile radius of one of 22 assessment centers and who were registered with the British National Health Service were invited by mail to participate (Allen et al., 2012; Sudlow et al., 2015). In total, 503,325 individuals agreed to participate in the baseline assessment (Allen et al., 2012). In 2014, the UK Biobank began inviting 100,000 individuals who participated in the baseline assessment for a follow-up imaging study (Alfaro-Almagro et al., 2018). As of January 14th, 2020, a total of 50,000 individuals have participated in the imaging study (UK Biobank, 2022), the data from whom were the focus of this investigation. All participants provided informed consent and were free to withdraw from the study at any time. The UK Biobank received ethics approval from the Northwest Multi-Centre Research Ethics Committee in 2011, which has since been renewed in 2016 and 2021 (for further details, see UK Biobank, 2021). The current investigation utilized the UK Biobank Resource through the NeuroHub platform under the Material Transfer Agreement held by McGill University (Application Reference Number: 45551). We joined the approved group ethics application that was handled by NeuroHub.

Participants

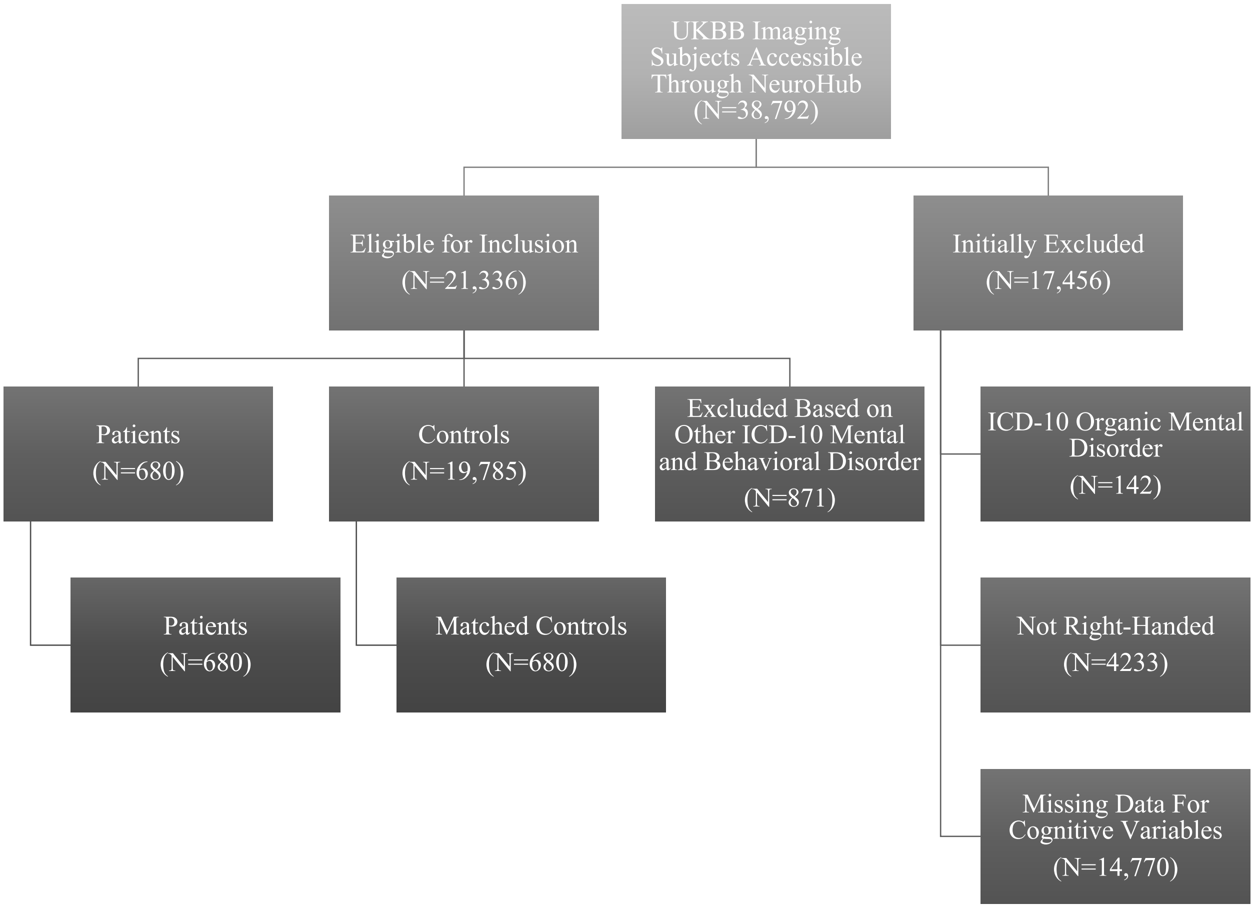
Data were archived on January 12th, 2022, using the NeuroHub platform and Compute Canada. Figure 1 depicts our sampling technique, which was designed to produce equally sized groups of patients and healthy controls. Diagnoses were recorded based on hospital inpatient records according to the International Classification of Diseases, 10th edition (ICD-10; World Health Organization, 2004). Individuals who were not right-handed, lacked data for any included cognitive tests, or who had ever been diagnosed with an organic mental disorder (ICD-10 codes: F00-F09) were excluded from analysis. Subsequently, individuals were split into patients, defined by a lifetime diagnosis of any mood, schizophrenia, schizotypal, or delusory condition (F20-F39); and controls, defined by the absence of any lifetime diagnosis of a mental disorder. Individuals who had another psychiatric disorder that was not co-morbid with a mood or psychotic disorder were also excluded from the analysis. Consistent with previous research (Baller et al., 2021), we obtained a matched group of controls based on age and sex. This was done using the Nearest Neighbors model in scikit-learn (Pedregosa et al., 2011) with the K-D Tree algorithm (Bentley, 1975).

Cognitive Measures

The cognitive tests used in the UK Biobank involved a cognitive battery administered through a touchscreen computer (Sudlow et al., 2015). All subjects who agreed to participate in the imaging follow-up were asked to complete four cognitive tests: the Prospective Memory Test, Pairs Matching Test (visual declarative memory), Fluid Intelligence Test (verbal and numerical reasoning), and Reaction Time Test. Eight additional tests were administered to subsamples: the Numeric Memory Test, Trail Making Test Parts A and B, Symbol Digit Substitution Test, Picture Vocabulary Test, Paired Associate Learning Test, Matrix Pattern Completion Test, and Tower Rearranging Test (Fawns-Ritchie & Deary, 2020). In many cases, these measures were adapted from more widely used cognitive tests, while others were devised specifically for use in the UK Biobank (Fawns-Ritchie & Deary, 2020). The present analysis included eight such measures designed to assess various aspects of memory, processing speed, and executive function. We excluded tests of verbal, numerical, and non-verbal reasoning (Fluid Intelligence Test, Matrix Pattern Completion Test), as well as crystalized ability (Picture Vocabulary Test), as impaired general intelligence is not considered a primary disturbance in depression (Marazziti et al., 2010). Below, we provide a brief description of each cognitive test used in the present analysis. Additional information regarding these tests, including an evaluation of their concurrent validity and test-retest reliability, is provided by Fawns-Ritchie and Deary (2020).

Figure 1

Sampling Procedure


UKBB Imaging Subjects Accessible Through NeuroHub (N=38,792)
Eligible for Inclusion (N=21,336)
Initially Excluded (N=17,456)
Patients (N=680)
Controls (N=19,785)
Excluded Based on Other ICD-10 Mental and Behavioral Disorder (N=871)
ICD-10 Organic Mental Disorder (N=142)
Patients (N=680)
Matched Controls (N=680)
Not Right-Handed (N=4233)
Missing Data For Cognitive Variables (N=14,770)

Prospective Memory Test

Before beginning the cognitive test battery, participants were told that they would later be shown four symbols and asked to touch the blue square, but they should touch the orange circle instead. After completing all other tests, participants were asked to touch the blue square (Fawns-Ritchie & Deary, 2020). Although the UK Biobank initially measured performance on a scale measuring the number of attempts required to recall the initial instruction, we quantified performance as a dichotomous variable depending on whether the participant remembered the original instruction, consistent with Fawns-Ritchie and Deary (2020).

Numeric Memory Test

The numeric memory test, which consisted of a backward digit span task, was used to assess working memory (Fawns-Ritchie & Deary, 2020). Participants were presented a two-digit number and then instructed to type the number in reverse order after a brief delay. If participants were correct, the length of the number increased by one on the next trial, otherwise it remained the same. This was repeated until the individual misremembered two trials with the same digit length, or until the person correctly remembered a 12-digit number (Fawns-Ritchie & Deary, 2020). The maximum number of digits accurately recalled was used to quantify performance.

Trail Making Test

Participants completed an adapted version of the Halstead-Reitan Trail Making Test (Reitan & Wolfson, 1985). The test was divided into two sections: Part A, which assessed processing speed, and Part B which assessed executive function (Bowie & Harvey, 2006). In Part A, subjects were presented with a screen displaying the numbers 1-25 in random order and directed to touch the digits in numerical order. In part B, the digits 1-13 and the letters A-L were arranged similarly randomly; subjects were told to alternate their selections between numerical and alphabetical order (Fawns-Ritchie & Deary, 2020). Performance was quantified for each part by the time required to complete the task successfully in seconds.

Pairs Matching Test

The Pairs Matching Test was used to evaluate visual declarative memory (Fawns-Ritchie & Deary, 2020). Participants were instructed to recall pairs of matching cards randomly arranged in a grid on the screen. Each participant performed two trials, the first of which included three pairs, while the second featured six pairs. Individuals who made less than three errors on the second trial were asked to complete a third trial, which included twelve pairs (Fawns-Ritchie & Deary, 2020). Following the reasoning of Fawns-Ritchie and Deary (2020), we used the number of attempts required on the second trial, as participants made relatively few errors on the first trial, and only a subsample completed the third trial.

Reaction Time Test

In the Reaction Time Test, participants were presented a series of card pairs and instructed to click a button to indicate whether the symbols on the cards were identical as quickly as possible. The mean time required to click the button on trials with matching cards was used to measure overall performance (Fawns-Ritchie & Deary, 2020).

Symbol Digit Substitution Test

Processing speed was assessed using the Symbol Digit Substitution Test (Fawns-Ritchie & Deary, 2020). Participants were presented a key that associated symbols to digits. Participants were shown symbols and instructed to name the corresponding digit for a duration of 60 seconds. The performance was based on the number of correct matches produced during the allotted time (Fawns-Ritchie & Deary, 2020).

Tower Rearranging Task

The Tower Rearranging Test was used to assess planning abilities, often described as a component of executive function (Fawns-Ritchie & Deary, 2020). On display A, participants were shown three pegs, on each of which three different color hoops were placed. On display B, the hoops were arranged differently. The participant's task was to work out the number of moves that it would take to make display A look like display B, which was between 1 and 6. The score was the number of items answered correctly in 3 minutes (Fawns-Ritchie & Deary, 2020).

Neuroimaging Markers

Neuroimaging data were acquired using Siemens Skyra 3T Scanners equipped with a 32-channel receiver head coil (Miller et al., 2016). The imaging assessments took place at at four locations throughout the United Kingdom, located in in Stockport, Newcastle-upon-Tyne, Reading, and Bristol (Littlejohns et al., 2020). Each location employed identical scanners, coils, software, and protocols (Littlejohns et al., 2020; Miller et al., 2016). The MRI protocol employed by the Biobank has been detailed by Alfaro-Almagro et al. (2018). Briefly, T1-weighted images were acquired using a 3D MPRAGE sequence (field of view: 208 × 256 × 256, voxel size: 1 × 1 × 1 mm, TR: 2000ms, TI: 800ms). Measures of cortical volume, surface area, and thickness were obtained using FreeSurfer (Fischl, 2012) for all 62 cortical regions of the Desikan-Killiany-Tourville (DKT) atlas (Klein & Tourville, 2012).

Statistical Analysis

Software

Analyses were carried out in Python 3.10 and our source code is freely available on GitHub (https://github.com/joshunrau/CognitiveSubtypes). All machine learning algorithms were implemented using scikit-learn v1.0.1, an open-source machine learning library written in Python that is designed to facilitate the application of both supervised and unsupervised machine learning algorithms to domain-specific tasks (Pedregosa et al., 2011).

Prior to beginning model development, a random subset of 25% of cases were chosen for the holdout set. As these cases were used for final validation, they were not included in any step of the model development process. This is essential so as to provide an objective metric of model performance (Vanderplas, 2016). Initially, we assessed the skewness of the distributions for all cognitive variables among patients and controls, which appeared similar in both groups. As scores for the Reaction Time, Symbol Digit, Prospective Memory, Numeric Memory, and Trail Making Tests appeared to be non-normally distributed, we applied a Yeo-Johnson power transformation (Yeo & Johnson, 2000) to enhance the properties of the distributions. This was based on the lambda function that maximized the loglikelihood in the training set, so as to not leak information regarding the distributional properties of the holdout set during the model development process (Müller & Guido, 2016). Scores on the Tower and Numeric Memory Tests appeared relatively normally distributed and hence no transformation was performed. Additionally, no transformation was considered for the Prospective Memory test, given that it was coded as a binary variable. Appendix A contains the distributions of all cognitive variables, both before and after any transformations. Subsequently, we calculated standardized scores for all cognitive variables using the mean and standard deviation of these variables within the training set.

Clustering

Next, we applied the k-means algorithm to acquire cluster labels for all cases based on the cognitive measures. In part due to its simplicity (Kodinariya & Makwana, 2013), the k-means algorithm remains one of the most widely adopted clustering methods in the literature

(Kodinariya & Makwana, 2013; Sinaga & Yang, 2020). In addition, the k-means algorithm is relatively simple to compute (Arthur & Vassilvitskii, 2006). For a given value for k, a set of k centroids (i.e., centers of clusters) are selected at random. After that, all values are assigned to a cluster based on the value of the closest centroid, and the centroids are updated to reflect the mean value assigned to them. This process continues until a point of equilibrium is reached (Müller & Guido, 2016, Chapter 3). Although it is widely recognized that there is no one-size-fits-all method for determining the optimal number of clusters (Burkov, 2019, Chapter 9; Kodinariya & Makwana, 2013), the average silhouette score is considered a balanced approach for determining the optimal value (Kodinariya & Makwana, 2013). The silhouette score computes the compactness of a cluster; higher is better, with a perfect score of 1 (Rousseeuw, 1987). Recent evidence continues to suggest that the silhouette score is an effective metric for determining the number of clusters (Shahapure & Nicholas, 2020). As an additional metric, we also examined the Calinski-Harabasz score, which is defined as the ratio of the between- and within-cluster dispersion (Caliński & Harabasz, 1974). We assessed model performance with two through six clusters, which we judged to be the highest number of clusters of any practical utility.

Classification

For background, any supervised learning model has two sets of parameters: model parameters, which are initialized and updated during the data learning process; and hyper-parameters, which define the model's architecture and therefore must be set prior to model training (Yang & Shami, 2020). To develop a model for cluster prediction, we evaluated three supervised learning algorithms: the K-Neighbours Classifier, the Ridge Classifier, and the Random Forest Classifier. Estimating the optimal set of hyperparameters is an essential part of building a machine learning model (Yang & Shami, 2020). Grid search is the simplest method for determining the optimal set of hyperparameters; it entails exhaustively searching among all possible combinations of candidates values from a given parameter space (Yang & Shami, 2020). While this method guarantees that the optimal parameters are chosen with respect to a given metric, it by definition becomes exponentially inefficient as the parameter space grows larger. Alternatively, Bayesian optimization provides a more efficient method of selecting hyperparameters. Rather than treating each hyperparameter configuration independently, values from previous search iterations are used to guide the parameters selected for the current iteration (Yang & Shami, 2020). For the present analyses, we used Bayesian optimization to optimize the hyperparameters for each model using the scikit-optimize library (Head et al., 2021). For the k-neighbours classifier, these parameters included the number of neighbors, the weight function used in prediction, and the distance metric used; for the ridge classifier, this included the optimal value for alpha (i.e., the regularization strength); for the random forest classifier, these parameters included the number of decision trees, maximum depth, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, and the maximum number of features to consider when looking for the best split.

Model training was performed using stratified fivefold cross validation. Briefly, this involved splitting the training data into five groups of equal size, each containing representative proportions of all clusters. Then, training was performed in five iterations, with each group held out to evaluate performance during one iteration, which was based on the area under the receiver operating curve (AUC; for a review of cross-validation, see Burkov, 2019, Chapter 5). The average AUC score across iterations was used to guide the hyperparameter search. After determining the optimal set of hyperparameters for each of the three models, we compared their performance within the training set (i.e., performance during cross-validation) as well as within the holdout set. This is recommended to reduce the risk of overfitting, as may be indicated by substantially higher performance in the training set than the holdout set (Burkov, 2019, Chapter 5).

Group Comparisons

Finally, we performed a series of statistical comparisons between groups for cognitive test scores, as well as age, sex, and diagnosis when applicable. The Kruskal–Wallis Test, which is robust to the assumption of homogeneity of variance (for an overview, see Ostertagova et al., 2014), was used to evaluate continuous outcome variables, whereas chi-squared tests were used to evaluate all categorical outcomes. For cognitive test scores that significantly differed between groups, Cohen's d effect sizes were calculated as the mean difference between groups divided by the pooled standard deviation (Cohen, 1988). In all cases, statistical significance was adjusted for multiple comparisons using the Bonferroni correction.

Results

Characteristics of the Sample

A total of 1360 individuals were included in the sample. Participants were predominantly female (64%) and were on average 62 years old at the time of imaging follow-up. The overwhelming majority of patients (n = 657, 96.6%) were diagnosed with only a mood disorder, while much smaller numbers had received a diagnosis of both an SSD and mood disorder (n = 12, 1.8%), or only an SSD (n = 11, 1.6%). Over 90% of patients had a lifetime diagnosis of an unspecified depressive episode (specific ICD-10 diagnoses are provided in Appendix B). Kruskal–Wallis tests did not indicate any statistically significant differences in any cognitive variables between diagnostic groups (see Appendix C). Regarding patients and controls, controls exhibited a trend towards slightly higher performance on all cognitive tasks. However, these differences were only statistically significant for the Trail Making Test Part A (Cohen's d = 0.19, *p* < .001) and Symbol Digit Test (Cohen's d = 0.41, p < .001). Additionally, controls demonstrated a strong trend toward superior performance on the pairs matching test, though this trend did not survive adjustment for multiple comparisons (p=0.007, significance: 0.05/8 cognitive tests = 0.006).

Table 1

Characteristics of Patients and Controls

	Controls (N=680)		Patients (N=680)			
	Mean/Percent	SD	Mean/Percent	SD	H	*p*
Age	62.28	7.47	62.28	7.47		
Female	64%		64%			
Reaction Time Test[a]	-0.06	0.98	0.00	1.04	0.59	0.441
Trail Making Test A[a]	**-0.07**	**1.00**	**0.12**	**1.03**	**14.66**	**<.001**
Trail Making Test B[a]	-0.05	0.97	0.06	1.05	1.15	0.284
Tower Test	0.02	0.96	-0.03	1.05	0.01	0.913
Symbol Digit Test	**0.2**	**1.00**	**-0.20**	**0.97**	**61.05**	**<.001**
Pairs Matching Test[a]	-0.05	0.96	0.07	1.05	7.36	0.007
Prospective Memory Test	0.01	0.99	-0.04	1.03	0.98	0.323
Numeric Memory Test	0.01	1.05	-0.05	0.98	1.97	0.16

Note. Rows depicted in bold represent those with significant group differences. As the standardization of cognitive test scores was accomplished using the mean and standard deviation of the training data, the mean and standard deviation of the overall sample may deviate from that of the standard normal distribution.

[a] Lower test scores indicate better performance

Cognitive Clusters

Figure 2 illustrates the Silhouette coefficients and Calinski-Harabasz indices for all clustering solutions evaluated. Both metrics indicated that binary clustering was the optimal solution. This solution was characterized by two relatively evenly sized clusters: a relatively high performing cluster (n=697, 51.3%) and a relatively low performing cluster (n=663, 48.7%). Table 2 summarises the cognitive and demographic characteristics of both clusters. Kruskal–Wallis tests revealed that the high-performing cluster was significantly younger than the low-performing cluster and outperformed them on all cognitive tasks (all p <. 001). The effect size was moderate for the Reaction Time Test (Cohen’s d = 0.5) and large for all other tests: Tower Test (Cohen’s d = 0.88), Pairs Matching Test (Cohen’s d = 0.96), Prospective Memory Test (Cohen’s d = 1.04), Numeric Memory Test (Cohen’s d = 1.10), Trail Making Test Part A (Cohen’s d = 1.40), Symbol Digit Test (Cohen’s d = 1.41), Trail Making Test Part B (Cohen’s d = 1.77). Groups did not significantly differ regarding sex or diagnosis (see also Appendix D for a violin plot of cluster performance by diagnosis).

Figure 2

Performance Metrics for Clustering Solutions

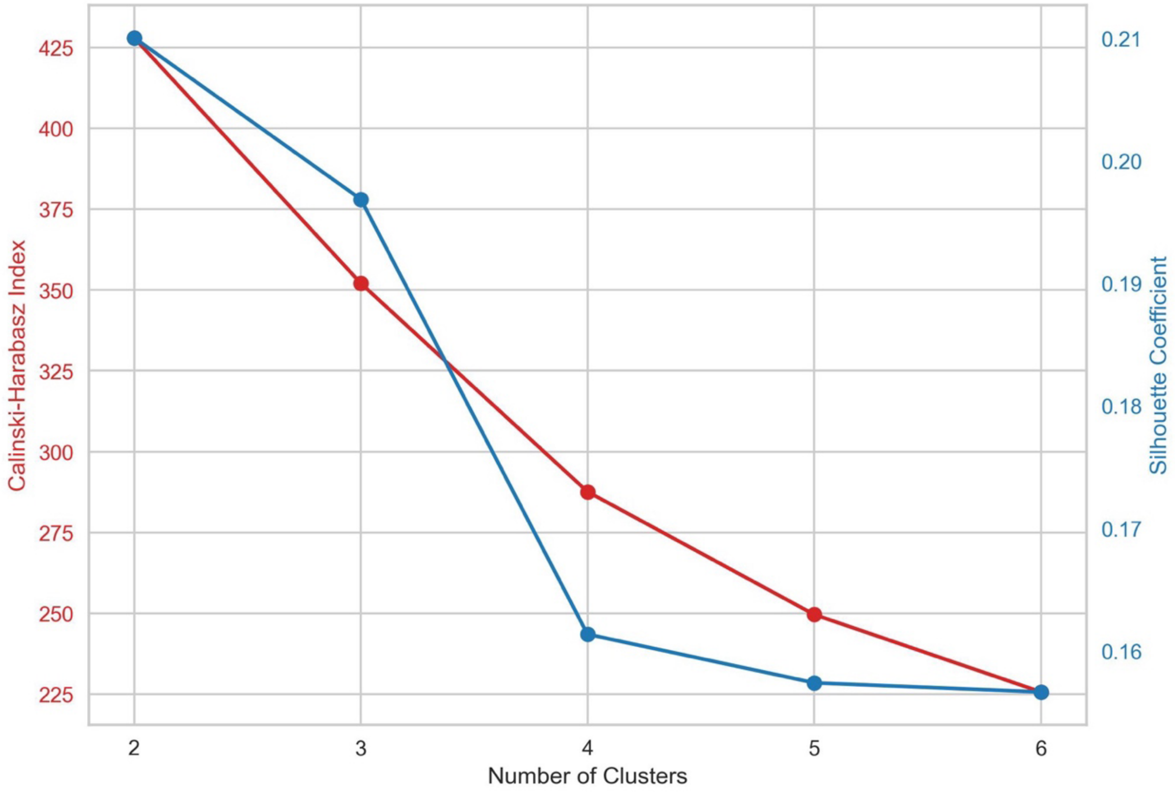


Table 2

Characteristics of Cognitive Clusters

	High (N= 697)		Low (N= 663)			
	Mean/Percent	SD	Mean/Percent	SD	Statistic	p
Age	59.07	6.67	65.61	6.77	H=259.95	<.001
Female	63%		65%		X2=.002	.968
Diagnosis					X2=.072	.995
…None	52.37%		47.51%			
…Only Mood Disorder	46.34%		50.38%			
…SSD and Mood Disorder	0.57%		1.06%			
…Only SSD	0.72%		1.06%			
Reaction Time Test[a]	**-0.27**	**0.99**	**0.22**	**0.98**	**H=74.54**	**<.001**
Trail Making Test A[a]	**-0.55**	**0.83**	**0.62**	**0.84**	**H=461.03**	**<.001**
Trail Making Test B[a]	**-0.64**	**0.78**	**0.69**	**0.72**	**H=626.16**	**<.001**
Tower Test	**0.4**	**0.96**	**-0.42**	**0.9**	**H=234.33**	**<.001**
Symbol Digit Test	**0.57**	**0.83**	**-0.59**	**0.81**	**H=499.62**	**<.001**
Pairs Matching Test[a]	**-0.43**	**0.82**	**0.44**	**0.99**	**H=244.02**	**<.001**
Prospective Memory Test	**0.41**	**0.49**	**-0.47**	**1.21**	**H=255.01**	**<.001**
Numeric Memory Test	**0.45**	**0.84**	**-0.52**	**0.93**	**H=310.88**	**<.001**

Note. SSD: Schizophrenia Spectrum Disorder; rows depicted in bold represent those with significant group differences.

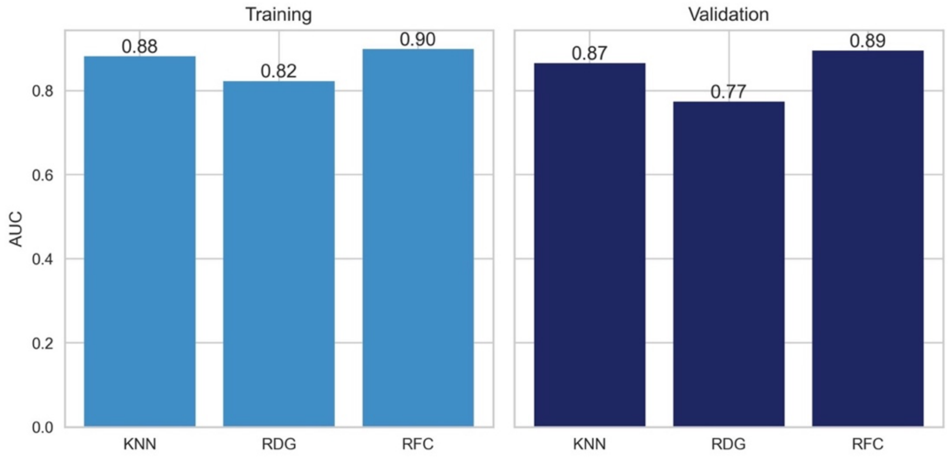
[a] Lower test scores indicate better performance

Predictive Model Selection

A total of 1020 subjects were used for model development, with 340 subjects reserved for final validation. As depicted in Figure 3, among the three models assessed, the Random Forest Classifier provided the highest performance in both the training set (AUC = .90) and validation set (AUC = .89). The Receiver Operating Curve for this model is presented in Figure 4. The overall accuracy of the model in the training set was 80%, while the model achieved 79% accuracy in the holdout set.

Figure 3

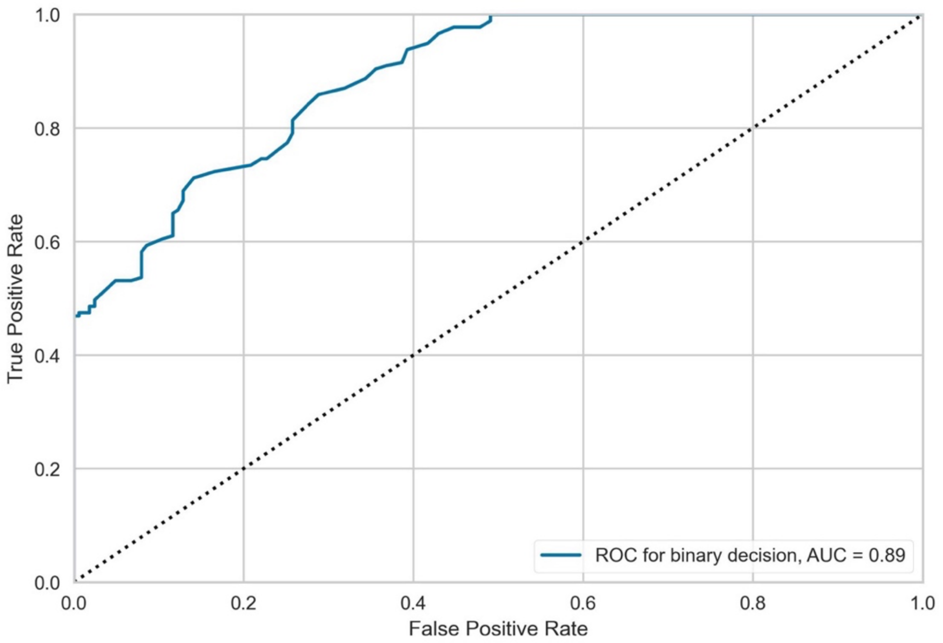
Area Under the Receiver Operating Curve for All Models Assessed



Note. AUC: Area Under the Curve, KNN: K-Nearest Neighbours Classifier, RDG: Ridge Classifier, RFC: Random Forest Classifier.

Figure 4

Receiver Operating Curve for Random Forest Classifier



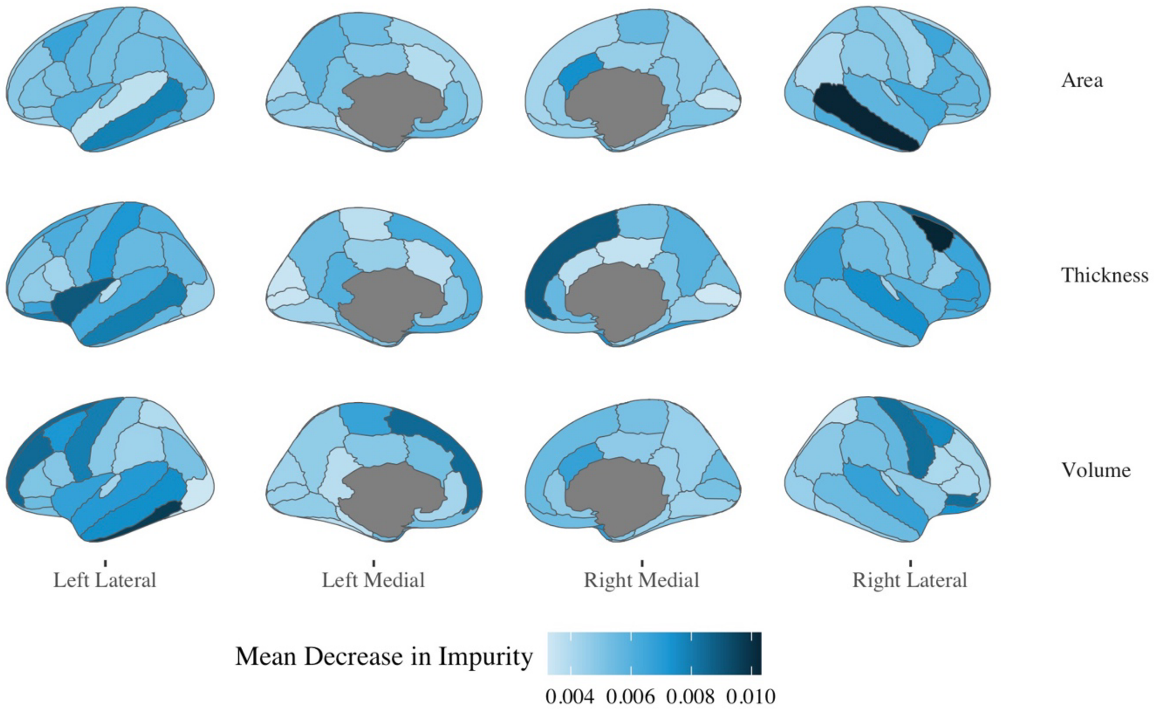
Note. ROC: Receiver Operating Curve, AUC: Area Under the Curve

Importance of Features in Predictive Model

Figure 5 depicts the importance of all features in the final model (the top 50 features are quantified in Appendix E). Feature importance is based on the mean decrease in impurity (i.e., the total decrease in node impurity for a given feature; see Louppe et al., 2013). Features with a relative importance greater than 80% (i.e., relative to the strongest feature component) included the right middle temporal gyrus area, right caudal middle frontal gyrus thickness, left inferior temporal gyrus volume, left insula thickness, right superior frontal gyrus thickness, left superior frontal gyrus volume, left rostral middle frontal gyrus volume, right precentral gyrus volume, and right pars orbitalis volume.

Figure 5

Feature Importance in Random Forest Model



Discussion

Summary of Findings

The present study had two principal objectives: (1) to delineate cognitive subtypes in a mixed sample of healthy controls and patients with psychotic and mood disorders; and (2) to predict these subtypes using measures of cortical morphometry. Our sample included 1360 subjects from the UK Biobank, of whom half were individuals with a lifetime diagnosis of a mood or psychotic disorder, while the other half were age and sex matched controls with no history of psychiatric illness. Using k-means clustering, we derived two roughly evenly sized cognitive subtypes: high performing (n=697, 51.3%) and low performing (n=663, 48.7%). To predict these clusters, we evaluated the performance of a variety of supervised learning algorithms, discovering that a Random Forest Classifier yielded the best performance; the model achieved an accuracy of 80% for training data (AUC = .90) and 79% for holdout data (AUC = .89). These findings demonstrate that cortical morphometry is capable of reliably differentiating cognitive subtypes in a mixed sample of patients with affective disorders and healthy controls.

Cognitive Performance of Patients and Controls

In our study, we observed relatively few differences between patients and controls regarding cognitive performance. Indeed, controls exhibited significantly better performance than patients on only two of eight cognitive tests: The Trail Making Test Part A and the Symbol Digit Test, both of which measure processing speed (Fawns-Ritchie & Deary, 2020). Overall, this is consistent with the results of a recent meta-analysis (Semkovska et al., 2019). However, while the effect size we observed for the Symbol Digit Test (Cohen’s d = 0.41) is comparable to that reported by the authors (Hedges’ g = 0.45), the same is not true for the Trail Making Test Part A;

we report a small effect size (Cohen's d = 0.19), whereas Semkovska et al. (2019) report a moderate effect size (Hedges' g = 0.54). This discrepancy might be explained by the fact that we performed a Yeo-Johnson power transformation, which decreases the relative influence of outlier values (Yeo & Johnson, 2000). Although this transformation was applied to both the Trail Making Test Part A and the Symbol Digit Test, there appeared to be substantially more outlier values for the Trail Making Test Part A (see Appendix A). If the probability that these outlier values belong to each cluster reflects the overall trend in cognitive performance (i.e., a greater proportion of subjects with exceptionally low scores on a given test belong to the low performing cluster, and vice versa), then it is reasonable to assume that reducing the size of outliers would reduce the magnitude of differences between clusters, and that this effect would be greater as the number of outliers increased. Therefore, the effect size for the Trail Making Test Part A would be reduced in our study.

Patients and controls did not significantly differ on six cognitive tests: the Reaction Time Test, Trail Making Test Part B, Tower Test, Pairs Matching Test, Prospective Memory Test, and Numeric Memory Test. It is not particularly surprising that patients did not differ from controls on the Reaction Time Test. Although some reports have indicated significant reaction time deficits in MDD patients, effect sizes have been small (Semkovska et al., 2019) and other studies have failed to find any difference between MDD patients and controls regarding reaction time (Rock et al., 2014). On the other hand, given that impaired executive functioning has been observed even in patients with remitted depression (Rock et al., 2014; Semkovska et al., 2019), it is somewhat surprising that we did not observe any significant differences between patients regarding the Trail Making Test Part B and Tower Test. However, there are several potential explanations for these findings. Given that the distribution of scores on the Trail Making Test

Part B was comparable to that on the Trail Making Test Part A, the Yeo-Johnson power transformation likely had a similar effect of reducing the magnitude of differences between patients and controls. Regarding the Tower Test, the ability to discriminate cognitive performance may have been impacted by its relatively poor psychometric properties. Indeed, the Tower Test is only moderately correlated (r = 0.40) with the D-KEFS Tower Test, the reference test it was based on (Fawns-Ritchie & Deary, 2020). Findings regarding memory performance are likely explained by two factors: the diagnoses present in our sample and the diagnostic criteria applied. Over 90% of patients in our sample were diagnosed with major depression and to be considered a patient, a subject had to have a lifetime diagnosis of a disorder, rather than a current diagnosis. It is well-established that a relatively large percentage of MDD patients experience complete recovery (Ahern & Semkovska, 2017). Therefore, it is possible that a large proportion of our patient sample were actually remitted patients. Given the inconclusive evidence for memory deficits in remitted MDD patients (Rock et al., 2014), it is therefore not particularly surprising that we observed no difference in memory performance between patients and controls.

Cognitive Clusters

Our discovery of a two-cluster solution contradicted our expectations, which were based on the findings of Baller et al. (2021). To our knowledge, this is the only study that examined subtypes defined primarily by cognitive performance across multiple domains in a sample with a large number of MDD patients. There are a number of potential explanations for this divergence. First, Baller et al. (2021) examined clusters specifically among individuals with a history of depression, whereas our analyses included healthy controls. Second, the UK Biobank is an older cohort, whereas Baller et al. (2021) conducted their analyses in a sample of youth with MDD.

Indeed, while the average age in our sample was 62, the average age in their sample was 16. Performance on most cognitive test is widely understood to be inversely correlated with age, as has recently been demonstrated in the UK Biobank (Fawns-Ritchie & Deary, 2020). Additionally, it is also unsurprising that our finding of a two cluster solution is not in line with the five subgroups found by Pelin et al. (2021). Although they used a sample similar to ours (i.e., significant proportions of both healthy controls and MDD patients), they derived subgroups based on 57 variables across a range of different domains beyond cognition. Thus, the subgroups derived likely reflected differences in a range of variables beyond cognition.

In contrast to the non-significant differences between clusters regarding diagnoses, we found that clusters significantly differed in age; the high performing cluster was significantly younger than the low performing cluster (Mean Age: 59.07 vs 65.61). This may be a major factor driving differences between groups. In the present study, the largest differences between clusters were observed for the Trail Making Test Part A, Symbol Digit Test, and Trail Making Test Part B. This is consistent with the work by Fawns-Ritchie and Deary (2020), who examined the correlations between performance on UK Biobank cognitive tests and age. They found the largest correlations were the Symbol Digit Test (r = -0.60), Trail Making Test Part A (r = 0.58), and Trail Making Test Part B (r = 0.57). Hence, clusters may reflect age-related changes in cortical structure.

Predictive Model

Our supervised learning model (random forest) achieved 79% accuracy in the holdout data. Initially, we reasoned that we would likely experience poorer performance compared to Gould et al. (2014), who reported an accuracy rate of 71%, due to the inclusion of a significant number of non-psychotic patients in our sample. However, the increased accuracy rate we observed in our study may be due to our larger sample size. Gould et al. (2014) included only 249 patients in their analyses, whereas we included 1380 individuals. It is well-known that larger amounts of training data tend to increase model performance (Müller & Guido, 2016; Vanderplas, 2016). In addition, unlike Gould et al. (2014), who did not evaluate model performance on a holdout set (i.e., the authors used only cross-validation on the training set), we provide an unbiased estimate of final model performance. Therefore, our model's high performance on holdout data is promising as it indicates good generalizability of our findings.

Although the estimates of feature importance we calculated for the Random Forest Classifier are not as straightforward to interpret as those from other models (e.g., linear regression), there are several brain areas of notable importance in our final model. In particular, right precentral gyrus volume was one of the most important features in our random forest model. This is similar to several previous findings in the literature. Alonso-Lana et al. (2016) found that compared to healthy controls, BD patients display reduced right precentral gyrus volume. In a mixed sample of SSD and BD patients, Shepherd et al. (2015) defined two subgroups they identify as executively spared or executively deficit. The executively deficit subgroup showed reduced grey matter volume in the right precentral gyrus compared to the executively spared subtype. Therefore, our findings contribute to a growing literature demonstrating that reduced right precentral gyrus volume may discriminate poorer cognitive performance. In addition, a number of the most important features were also those which are associated with age-related changes. Notably, we identified right superior frontal gyrus thickness as an important feature in the model. It is established across studies that cortical thinning in the superior frontal gyrus displays strong age effects (Fjell et al., 2009). Age effects have also been consistently noted in the middle temporal gyrus (Fjell et al., 2009). In our model, area of the right middle temporal gyrus was the most important feature, which lends credence to the hypothesis that many of the differences between clusters may reflect aging-related structural changes.

Limitations

There are several important limitations of this study that should be considered. First, as our research question was developed prior to looking at the data, our study included a relatively small number of SSD patients. Given that cognitive impairments are most severe in this population (Millan et al., 2012), this limits the potential clinical utility of our work in addressing the needs of this high-risk population. Although, it is important to note that we also provide much needed research regarding cognitive deficits in an understudied population (i.e., MDD patients). Second, as assessment of social cognition was not available in the UK Biobank, we did not assess theory of mind, which might discriminate SSDs from affective psychoses (Bora et al., 2016). Also, we did not control for age, which is well-known to be associated with general cortical thinning (Fjell et al., 2009). Finally, we did not make any statistical comparisons of brain regions between clusters. Although this was not the purpose of our study, this precludes making any definitive statements regarding cortical differences between clusters.

Future Directions

Future studies should consider controlling for age when developing supervised learning models. This could yield particular benefits when considering cognitive clusters in exclusively clinical populations, as the importance of features might reflect potential biological subtypes more accurately. In addition, while currently somewhat limited by computational feasibility, researchers should examine whether more complex features could potentially provide a better fit for the data. For example, Yang et al. (2021) recently established two subtypes of MDD patients based on structural covariance networks (SCNs); these groups significantly differed with respect to cognitive performance. Potentially, future research could examine whether this result can be confirmed from the reverse direction (e.g., patterns of SCNs associated with cognitive subtypes). Finally, researchers who are adept at algorithm development could base clustering on the ability of later supervised models to predict these clusters in a unified model based on a backpropagation-like system. This could allow for the discovery of the cognitive subtypes that best load onto neurological markers.

Acknowledgements

I would like to express my gratitude to Dr. Martin Lepage for his guidance and supervision over the past two years. Under his supervision, I have had the opportunity to refine my academic interests, present work at academic conferences, and become capable of conducting independent research. None of this would have been possible without his unwavering support. I would also like to thank Dr. Katie Lavigne for all the time she has spent helping me with this project, without whom this project would also have not been possible. She has provided incredible support and guidance as I developed this study, and thoughtful and relevant insights in interpreting the results.

Statement of Contribution

J.U. was principally responsible for conceiving of the general study design, performing background research, developing the methodology, writing source code, running analyses, and interpreting the results. K.L. provided guidance regarding all of the above responsibilities. M.L. also provided guidance regarding all of the above responsibilities, with the exception of writing source code.

References

Ahern, E., & Semkovska, M. (2017). Cognitive functioning in the first-episode of major depressive disorder: A systematic review and meta-analysis. *Neuropsychology*, *31*(1), 52-72. https://doi.org/10.1037/neu0000319

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., . . . Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*, *166*, 400-424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P., Pell, J., Sprosen, T., & Collins, R. (2012). UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, *1*(3), 123-126. https://doi.org/https://doi.org/10.1016/j.hlpt.2012.07.003

Alonso-Lana, S., Goikolea, J. M., Bonnin, C. M., Sarró, S., Segura, B., Amann, B. L., Monté, G. C., Moro, N., Fernandez-Corcuera, P., & Maristany, T. (2016). Structural and functional brain correlates of cognitive impairment in euthymic patients with bipolar disorder. *PLoS One*, *11*(7), e0158867.

Arthur, D., & Vassilvitskii, S. (2006). *k-means++: The advantages of careful seeding*.

Baller, E. B., Kaczkurkin, A. N., Sotiras, A., Adebimpe, A., Bassett, D. S., Calkins, M. E., Chand, G. B., Cui, Z., Gur, R. E., Gur, R. C., Linn, K. A., Moore, T. M., Roalf, D. R., Varol, E., Wolf, D. H., Xia, C. H., Davatzikos, C., & Satterthwaite, T. D. (2021). Neurocognitive and functional heterogeneity in depressed youth. *Neuropsychopharmacology*, *46*(4), 783-790. https://doi.org/10.1038/s41386-020-00871-w

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, *18*(9), 509-517.

Bleuler, E. (1911). *Dementia Praecox, or Group of Schizophrenias*. (Dementia Praecox, oder Gruppe der Schizophrenien)

Bora, E., Veznedaroglu, B., & Vahip, S. (2016). Theory of mind and executive functions in schizophrenia and bipolar disorder: A cross-diagnostic latent class analysis for identification of neuropsychological subtypes. *Schizophr Res*, *176*(2-3), 500-505. https://doi.org/10.1016/j.schres.2016.06.007

Bora, E., Yucel, M., & Pantelis, C. (2009). Cognitive functioning in schizophrenia, schizoaffective disorder and affective psychoses: meta-analytic study. *The British Journal of Psychiatry*, *195*(6), 475-482.

Bora, E., Yücel, M., & Pantelis, C. (2010). Neurocognitive markers of psychosis in bipolar disorder: A meta-analytic study. *Journal of affective disorders*, *127*(1), 1-9. https://doi.org/https://doi.org/10.1016/j.jad.2010.02.117

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the Trail Making Test. *Nature protocols*, *1*(5), 2277-2281.

Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov. https://books.google.ca/books?id=0jbxwQEACAAJ

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1-27.

Cobia, D. J., Csernansky, J. G., & Wang, L. (2011). Cortical thickness in neuropsychologically near-normal schizophrenia. *Schizophrenia research*, *133*(1-3), 68-76.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates. *Hillsdale, NJ*, 20-26.

East-Richard, C., R. -Mercier, A., Nadeau, D., & Cellard, C. (2020). Transdiagnostic neurocognitive deficits in psychiatry: A review of meta-analyses. *Canadian Psychology/Psychologie canadienne*, *61*(3), 190-214. https://doi.org/10.1037/cap0000196

Fawns-Ritchie, C., & Deary, I. J. (2020). Reliability and validity of the UK Biobank cognitive tests. *PLoS One*, *15*(4), e0231627. https://doi.org/10.1371/journal.pone.0231627

Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*(2), 774-781.

Fjell, A. M., Westlye, L. T., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., Agartz, I., Salat, D. H., Greve, D. N., & Fischl, B. (2009). High consistency of regional cortical thinning in aging across multiple samples. *Cerebral cortex*, *19*(9), 2001-2012.

Geisler, D., Walton, E., Naylor, M., Roessner, V., Lim, K. O., Schulz, S. C., Gollub, R. L., Calhoun, V. D., Sponheim, S. R., & Ehrlich, S. (2015). Brain structure and function correlates of cognitive subtypes in schizophrenia. *Psychiatry Research: Neuroimaging*, *234*(1), 74-83.

Gould, I. C., Shepherd, A. M., Laurens, K. R., Cairns, M. J., Carr, V. J., & Green, M. J. (2014). Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: a support vector machine learning approach. *NeuroImage: Clinical*, *6*, 229-236.

Green, M. F., & Harvey, P. D. (2014). Cognition in schizophrenia: Past, present, and future. *Schizophr Res Cogn*, *1*(1), e1-e9. https://doi.org/10.1016/j.scog.2014.02.001

Green, M. J., Girshkin, L., Kremerskothen, K., Watkeys, O., & Quide, Y. (2020). A Systematic Review of Studies Reporting Data-Driven Cognitive Subtypes across the Psychosis Spectrum. *Neuropsychol Rev*, *30*(4), 446-460. https://doi.org/10.1007/s11065-019-09422-7

Head, T., Kumar, M., Nahrstaedt, H., Louppe, G., & Shcherbatyi, I. (2021). scikit-optimize/scikit-optimize (v0.9.0). *Zenodo*. https://doi.org/https://doi.org/10.5281/zenodo.5565057

Heinrichs, R. W., & Zakzanis, K. K. (1998). Neurocognitive deficit in schizophrenia: A quantitative review of the evidence. *Neuropsychology*, *12*(3), 426-445. https://doi.org/10.1037/0894-4105.12.3.426

Hill, S. K., Reilly, J. L., Keefe, R. S., Gold, J. M., Bishop, J. R., Gershon, E. S., Tamminga, C. A., Pearlson, G. D., Keshavan, M. S., & Sweeney, J. A. (2013). Neuropsychological impairments in schizophrenia and psychotic bipolar disorder: findings from the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP) study. *American Journal of Psychiatry*, *170*(11), 1275-1284.

Karantonis, J. A., Carruthers, S. P., Burdick, K. E., Pantelis, C., Green, M., Rossell, S. L., Hughes, M. E., Cropley, V., & Van Rheenen, T. E. (2022). Brain Morphological Characteristics of Cognitive Subgroups of Schizophrenia-Spectrum Disorders and Bipolar Disorder: A Systematic Review with Narrative Synthesis. *Neuropsychology Review*. https://doi.org/10.1007/s11065-021-09533-0

Klein, A., & Tourville, J. (2012). 101 Labeled Brain Images and a Consistent Human Cortical Labeling Protocol [Original Research]. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00171

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.

Krabbendam, L., Arts, B., van Os, J., & Aleman, A. (2005). Cognitive functioning in patients with schizophrenia and bipolar disorder: A quantitative review. *Schizophrenia research*, *80*(2), 137-149. https://doi.org/https://doi.org/10.1016/j.schres.2005.08.004

Kraepelin, E. (1919). *Dementia Praecox and Paraphrenia*.

Kraepelin, E. (1921). Manic Depressive Insanity and Paranoia.

Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., Bell, J. D., Boultwood, C., Collins, R., & Conroy, M. C. (2020). The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications*, *11*(1), 1-12.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems*, *26*.

Marazziti, D., Consoli, G., Picchetti, M., Carlini, M., & Faravelli, L. (2010). Cognitive impairment in major depression. *European journal of pharmacology*, *626*(1), 83-86.

McAllister, T. W. (1981). Cognitive functioning in the affective disorders. *Comprehensive Psychiatry*, *22*(6), 572-586. https://doi.org/10.1016/0010-440X(81)90006-7

Millan, M. J., Agid, Y., Brune, M., Bullmore, E. T., Carter, C. S., Clayton, N. S., Connor, R., Davis, S., Deakin, B., DeRubeis, R. J., Dubois, B., Geyer, M. A., Goodwin, G. M., Gorwood, P., Jay, T. M., Joels, M., Mansuy, I. M., Meyer-Lindenberg, A., Murphy, D., . . . Young, L. J. (2012). Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy. *Nat Rev Drug Discov*, *11*(2), 141-168. https://doi.org/10.1038/nrd3628

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., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., . . . Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*, *19*(11), 1523-1536. https://doi.org/10.1038/nn.4393

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning With Python: A Guide for Data Scientists*. O'Reilly Media, Inc.

Ostertagova, E., Ostertag, O., & Kováč, J. (2014). Methodology and application of the Kruskal-Wallis test. Applied Mechanics and Materials,

Pan, Z., Park, C., Brietzke, E., Zuckerman, H., Rong, C., Mansur, R. B., Fus, D., Subramaniapillai, M., Lee, Y., & McIntyre, R. S. (2019). Cognitive impairment in major depressive disorder. *Cns Spectrums*, *24*(1), 22-29.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pelin, H., Ising, M., Stein, F., Meinert, S., Meller, T., Brosch, K., Winter, N. R., Krug, A., Leenings, R., & Lemke, H. (2021). Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning. *Neuropsychopharmacology*, *46*(11), 1895-1905.

Perini, G., Ramusino, M. C., Sinforiani, E., Bernini, S., Petrachi, R., & Costa, A. (2019). Cognitive impairment in depression: recent advances and novel treatments. *Neuropsychiatric disease and treatment*, *15*, 1249.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation* (Vol. 4). Reitan Neuropsychology.

Rock, P. L., Roiser, J. P., Riedel, W. J., & Blackwell, A. D. (2014). Cognitive impairment in depression: a systematic review and meta-analysis. *Psychol Med*, *44*(10), 2029-2040. https://doi.org/10.1017/S0033291713002535

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. https://doi.org/10.1147/rd.33.0210

Semkovska, M., Quinlivan, L., O'Grady, T., Johnson, R., Collins, A., O'Connor, J., Knittle, H., Ahern, E., & Gload, T. (2019). Cognitive function following a major depressive episode: a systematic review and meta-analysis. *The Lancet Psychiatry*, *6*(10), 851-861.

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA),

Shepherd, A. M., Quidé, Y., Laurens, K. R., O'Reilly, N., Rowland, J. E., Mitchell, P. B., Carr, V. J., & Green, M. J. (2015). Shared intermediate phenotypes for schizophrenia and bipolar disorder: neuroanatomical features of subtypes distinguished by executive dysfunction. *Journal of Psychiatry and Neuroscience*, *40*(1), 58-68.

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716-80727.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

UK Biobank. (2021). *UK Biobank research ethics approval*. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics

UK Biobank. (2022). *World's largest imaging study scans 50,000th participant*. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/world-s-largest-imaging-study-scans-50-000th-participant

Van Rheenen, T. E., Cropley, V., Zalesky, A., Bousman, C., Wells, R., Bruggemann, J., Sundram, S., Weinberg, D., Lenroot, R. K., Pereira, A., Shannon Weickert, C., Weickert, T. W., & Pantelis, C. (2018). Widespread Volumetric Reductions in Schizophrenia and Schizoaffective Patients Displaying Compromised Cognitive Abilities. *Schizophr Bull*, *44*(3), 560-574. https://doi.org/10.1093/schbul/sbx109

Van Rheenen, T. E., Lewandowski, K. E., Tan, E. J., Ospina, L. H., Ongur, D., Neill, E., Gurvich, C., Pantelis, C., Malhotra, A. K., Rossell, S. L., & Burdick, K. E. (2017). Characterizing cognitive heterogeneity on the schizophrenia-bipolar disorder spectrum. *Psychol Med*, *47*(10), 1848-1864. https://doi.org/10.1017/S0033291717000307

Vanderplas, J. (2016). Machine Learning. In *Python Data Science Handbook* (pp. 331-515). O'Reilly Media, Inc.

Vöhringer, P., Barroilhet, S., Amerio, A., Reale, M., Vergne, D., Alvear, K., & Ghaemi, S. (2013). Cognitive Impairment in Bipolar Disorder and Schizophrenia: A Systematic Review [Review]. *Frontiers in Psychiatry*, *4*. https://doi.org/10.3389/fpsyt.2013.00087

World Health Organization. (2004). *ICD-10 : international statistical classification of diseases and related health problems : tenth revision* (2nd ed.). World Health Organization. https://apps.who.int/iris/handle/10665/42980

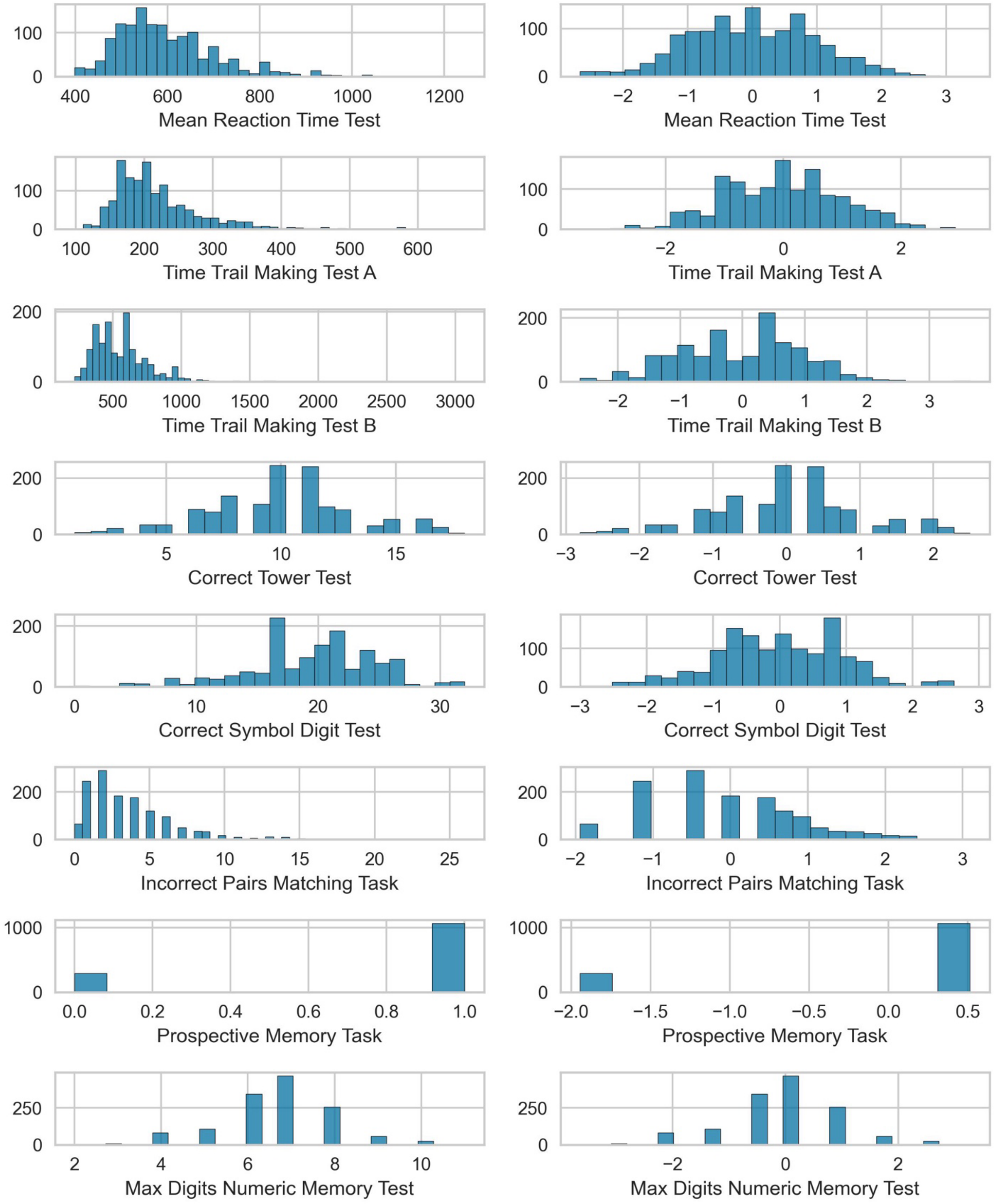
Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316.

Yang, X., Qi, S., Wang, M., Calhoun, V. D., Sui, J., Li, T., & Ma, X. (2021). Subtypes of depression characterized by different cognitive decline and brain activity alterations. *J Psychiatr Res*, *138*, 413-419. https://doi.org/10.1016/j.jpsychires.2021.04.023

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954-959. https://doi.org/10.1093/biomet/87.4.954

Appendix A

Distributions of Cognitive Variables



Appendix B

ICD-10 Diagnoses

Diagnosis	N	%	Age (Mean)	Female (%)
F20.0 Paranoid schizophrenia	3	0.44	58.67	33.33
F20.2 Catatonic schizophrenia	1	0.15	55	100
F20.9 Schizophrenia, unspecified	9	1.32	64.22	22.22
F22.0 Delusional disorder	7	1.03	61.71	28.57
F23.9 Acute and transient psychotic disorder, unspecified	3	0.44	63.67	0
F25.0 Schizoaffective disorder, manic type	1	0.15	53	0
F25.9 Schizoaffective disorder, unspecified	2	0.29	61	50
F29 Unspecified nonorganic psychosis	4	0.59	61.75	75
F30.0 Hypomania	3	0.44	61.67	33.33
F30.2 Mania with psychotic symptoms	2	0.29	60	50
F30.9 Manic episode, unspecified	3	0.44	63	66.67
F31.0 Bipolar affective disorder, current episode hypomanic	2	0.29	58.5	100
F31.1 Bipolar affective disorder, current episode manic without psychotic symptoms	4	0.59	58.25	75
F31.2 Bipolar affective disorder, current episode manic with psychotic symptoms	4	0.59	57.25	50
F31.4 Bipolar affective disorder, current episode severe depression without psychotic symptoms	1	0.15	58	100
F31.6 Bipolar affective disorder, current episode mixed	2	0.29	65.5	50
F31.9 Bipolar affective disorder, unspecified	26	3.82	61.54	50
F32.0 Mild depressive episode	2	0.29	63	100
F32.1 Moderate depressive episode	7	1.03	63.14	28.57
F32.2 Severe depressive episode without psychotic symptoms	13	1.91	62.62	61.54
F32.3 Severe depressive episode with psychotic symptoms	5	0.74	59.4	20
F32.9 Depressive episode, unspecified	623	91.62	62.34	66.13
F33.1 Recurrent depressive disorder, current episode moderate	2	0.29	65	50
F33.2 Recurrent depressive disorder, current episode severe without psychotic symptoms	1	0.15	62	100
F33.3 Recurrent depressive disorder, current episode severe with psychotic symptoms	2	0.29	71	0
F33.9 Recurrent depressive disorder, unspecified	15	2.21	58.6	73.33
F34.1 Dysthymia	2	0.29	67	50
F39 Unspecified mood [affective] disorder	1	0.15	69	100

Note. Totals may be greater than the number of patients due to comorbidity

Appendix C

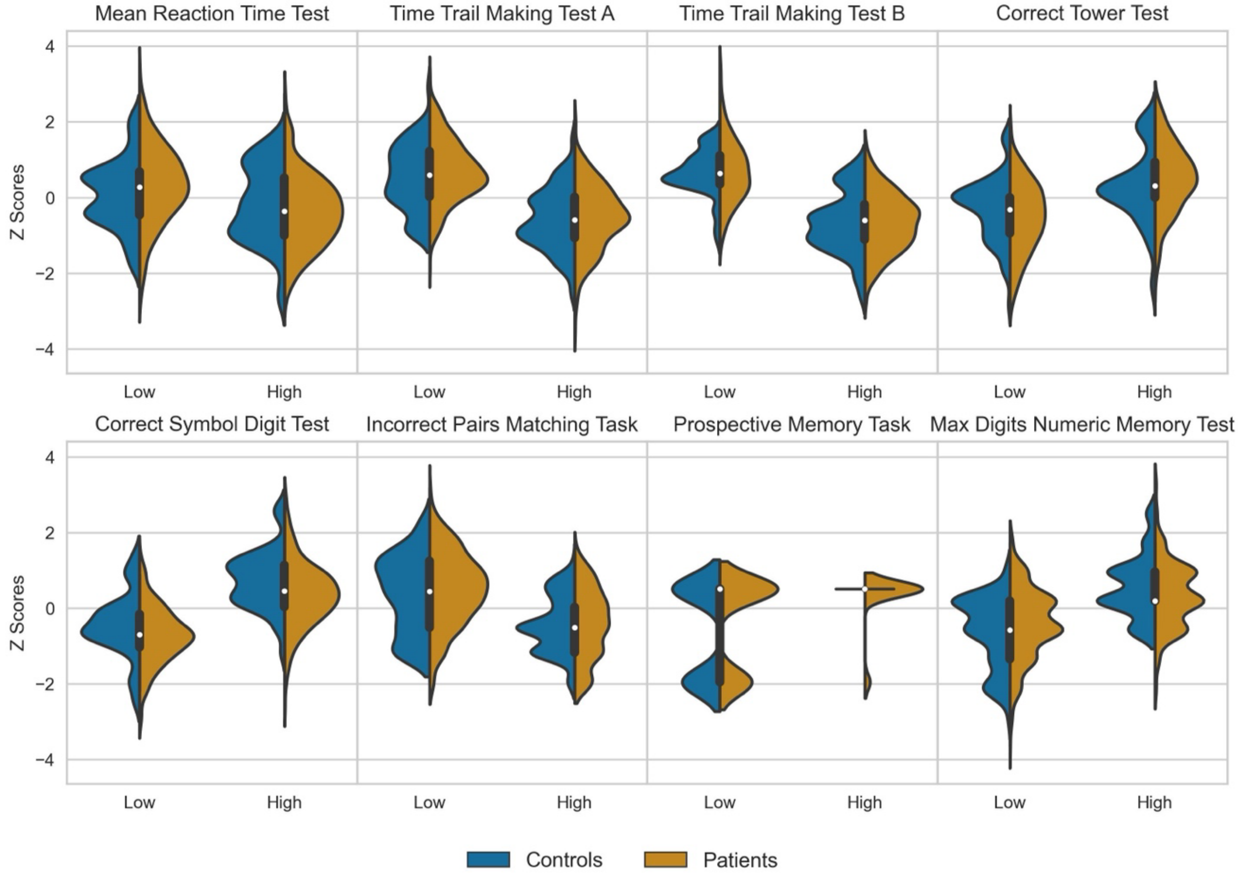
Performance Between Diagnostic Groups on Cognitive Tests

	SSD		SSD + MD		MD			
	M	SD	M	SD	M	SD	H	p
Reaction Time Test	0.34	1.05	-0.46	1.35	0	1.04	3.07	0.215
Trail Making Test Part A	0.32	1.15	0.19	0.75	0.11	1.03	0.72	0.698
Trail Making Test Part B	0.09	1.16	0.34	0.85	0.06	1.04	1.09	0.579
Tower Test	-0.4	1.22	0.21	1.31	-0.03	1.05	2.43	0.297
Symbol Digit Test	-0.32	0.94	-0.25	0.77	-0.2	0.97	0.2	0.904
Pairs Matching Test	0.35	0.86	0.01	1.33	0.05	1.04	1.05	0.591
Prospective Memory Test	0.06	1	0.3	0.71	-0.06	1.04	1.57	0.456
Numeric Memory Test	0.26	0.64	-0.07	0.68	-0.06	0.99	1.48	0.477

Note. M: Mean, SD: Standard Deviation, SSD: Schizophrenia Spectrum Disorder, MD: Mood Disorder

Appendix D

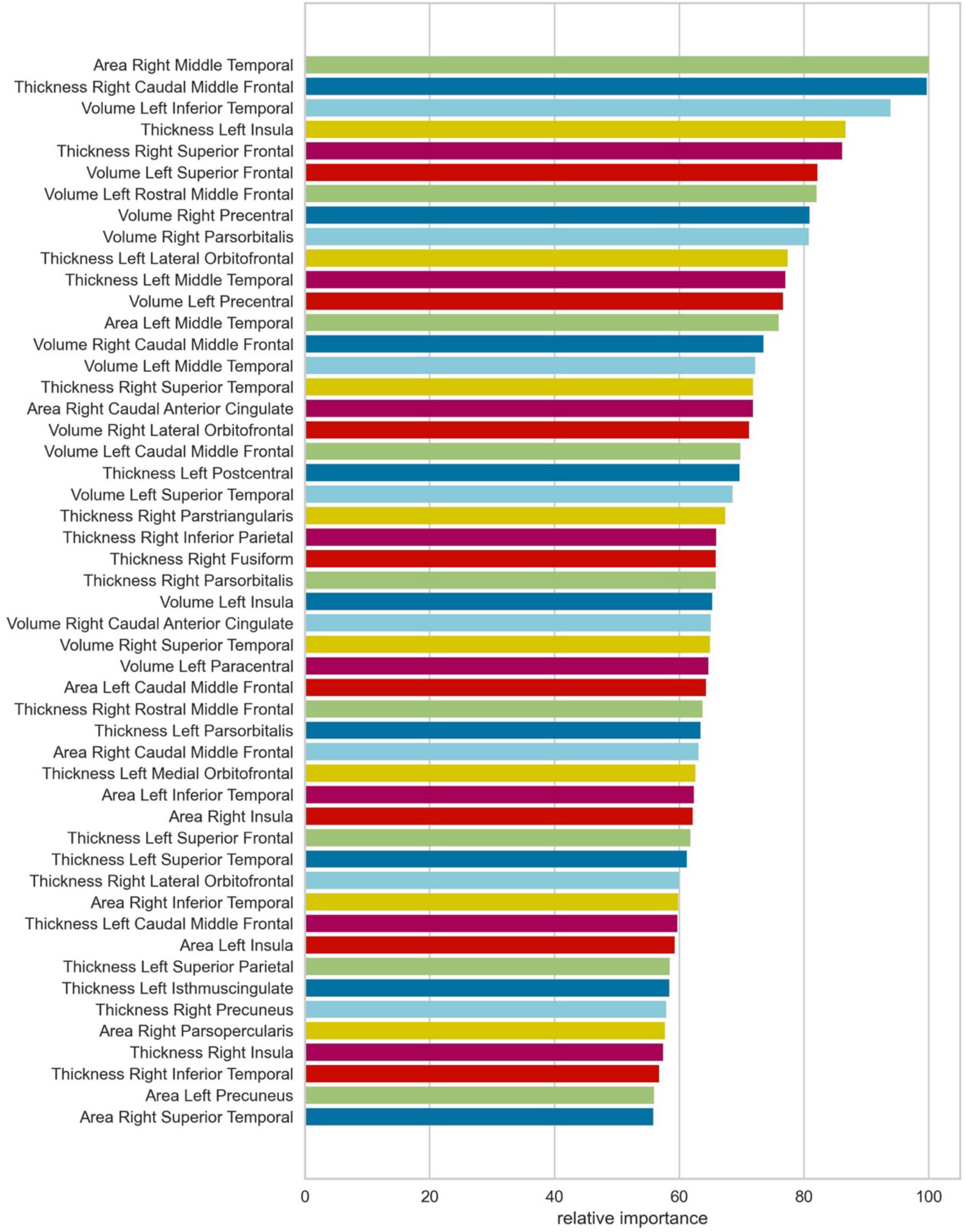
Violin Plot of Cognitive Performance



Note. High: High Performing Cluster, Low: Low Performing Cluster

Appendix E

Relative Importance of Top 50 Features in Random Forest Model



Note. Feature importance was based on the mean decrease in impurity (MDI)