Multimodal structural neuroimaging data unveil data-driven subtypes of treatment-resistant depression.

Federica Colombo\textsuperscript{1,2}, Federico Calesella\textsuperscript{1,2}, Beatrice Bravi\textsuperscript{1,2}, Lidia Fortaner-Uyà\textsuperscript{1,2}, Camilla Monopoli\textsuperscript{2}, Emma Tassi\textsuperscript{3}, Matteo Carminati\textsuperscript{1}, Raffaella Zanardi\textsuperscript{1,4}, Irene Bollettini\textsuperscript{2}, Sara Poletti\textsuperscript{1,2}, Paolo Brambilla\textsuperscript{3}, Alessandro Serretti\textsuperscript{3}, Eleonora Maggioni\textsuperscript{3}, Chiara Fabbri\textsuperscript{5,6}, Francesco Benedetti\textsuperscript{1,2}, Benedetta Vai\textsuperscript{1,2}

\textsuperscript{1} University Vita-Salute San Raffaele, Milano, Italy
\textsuperscript{2} Psychiatry and Clinical Psychobiology Unit, Division of Neuroscience, IRCCS San Raffaele Hospital, Milano, Italy
\textsuperscript{3} Department of Neurosciences and Mental Health, IRCCS Fondazione Ca’ Granda Ospedale Maggiore Policlinico, Milan, Italy
\textsuperscript{4} Mood Disorders Unit, Scientific Institute IRCCS San Raffaele Hospital, Milan, Italy
\textsuperscript{5} Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy.
\textsuperscript{6} Institute of Psychiatry, Psychology & Neuroscience, King’s College London, UK

To whom correspondence should be addressed:

Dr. Federica Colombo. M.Sc.
IRCCS Ospedale San Raffaele
Division of Neuroscience, Unit of psychiatry and clinical psychobiology
San Raffaele Turro, Via Stamira d’Ancona 20
Milano, Italy
Tel +39/02/26433156
Fax +39/02/26433265
E-mail: f.colombo8@studenti.unisr.it

Running title: neuroimaging biotypes of treatment-resistant depression

Keywords: clustering, major depressive disorder, treatment-resistant depression, machine learning, structural neuroimaging

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Background: An estimated 30% of Major Depressive Disorder (MDD) patients exhibit resistance to conventional antidepressant treatments. Identifying reliable biomarkers of treatment-resistant depression (TRD) represents a major goal of precision psychiatry, which is hampered by the clinical and biological heterogeneity underlying MDD.

Methods: To parse heterogeneity and uncover biologically-driven subtypes of MDD, we applied an unsupervised data-driven framework to stratify 102 MDD patients on their neuroimaging signature, including extracted measures of cortical thickness, grey matter volumes, and white matter fractional anisotropy. Our novel analytical pipeline integrated different machine learning algorithms to harmonize neuroimaging data, perform data dimensionality reduction, and provide a stability-based relative clustering validation. The obtained clusters were then characterized for TRD, history of childhood trauma and different profiles of depressive symptoms.

Results: Our results indicated two different clusters of patients, differentiable with 67% of accuracy: 1) one cluster (n=59) was associated with a higher proportion of TRD compared to the other, and higher scores of energy-related depressive symptoms, history of childhood abuse and emotional neglect; this cluster showed a widespread reduction in cortical thickness and volumes, along with fractional anisotropy in the right superior fronto-occipital fasciculus, stria terminalis, and corpus callosum; 2) the second cluster (n=43) was associated with cognitive and affective depressive symptoms and thicker cortices and wider volumes compared to the other.

Discussion: Our stratification of MDD patients based on structural neuroimaging identified clinically-relevant subgroups of TRD with specific symptomatic and childhood trauma profiles, which are informative for tailoring personalized and more effective interventions of treatment resistance.
Introduction

Major depressive disorder (MDD) represents one of the most common psychiatric disorders worldwide, with almost 17% of men and 25% of women experiencing at least one depressive episode during their lifetime (1). One of the most alarming issues contributing to the social and economic burden of the disorder is the uncertainty linked to clinical outcomes: despite the wide variability in treatment modalities, about 30% of patients do not achieve remission after antidepressant treatments (2) or relapse soon after, paving the way to treatment-resistant depression (TRD) (3). The lack of response to at least one antidepressant treatment is associated with more severe clinical outcomes, such as higher risk of suicide, more frequent relapses, psychiatric and medical comorbidities, health deterioration and functional impairment (4, 5). The prediction of clinical outcomes is further complicated by the absence of objective criteria for treatment choice, and several trials are often needed to find the optimal treatment for a patient (6). Indeed, there is a dire need for a “precision medicine” approach to MDD for the development of more effective treatments tailored on the individual profile.

The failure of achieving remission to at least two antidepressant treatments of adequate dose and duration is the most commonly used definition of TRD (7), and a recent general consensus on the clinical definition of TRD recommended future subgroup or stratified analyses on biological samples, because no biomarker has yet been validated to identify people with TRD (3). One of the main obstacles hampering the development of personalized treatment strategies is indeed the clinical and biological heterogeneity of MDD. Different clinical subtypes of clinical manifestations have been recognized according to established diagnostic criteria (8), and most of them have been associated with TRD (9). Although the heritability of MDD is estimated at 35-40% (10), approximately 67% of variance in MDD is explained by environmental factors (11). An example of such factors is the experience of childhood trauma, which is considered as one of the strongest predictors of poor antidepressant response (12-14). Compared to those not exposed to adverse childhood experiences, individuals with an history of childhood trauma typically have an earlier onset of MDD, more deleterious course, lower rates of remissions and response to treatments (15, 16). However, the extent to which the aetiology of TRD is shared with treatment-responsive depression is still unclear. Therefore, the delineation of objective biomarkers of TRD is of the utmost importance to define novel treatments targeting the underlying neurobiology. In terms of neural correlates, structural neuroimaging studies indicate that volumetric changes in fronto-striatal and...
hippocampal volumes, together with reduced structural integrity in the corpus callosum and superior longitudinal fasciculus, differentiate TRD from treatment-responsive MDD and healthy controls (17, 18). Nevertheless, most of these findings failed to be consistently replicated, possibly due to the heterogeneity in the assessment of treatment resistance and in statistical methods (19).

As an attempt to uncover subgroups of psychiatric conditions, several studies employed unsupervised machine learning approaches in order to stratify patients based on shared characteristics (20-22). The most common approach is to apply unsupervised clustering algorithms on clinical data. While these studies suggest that specific symptoms-based clusters could be possibly linked to different effectiveness of antidepressant treatments (23-25), the analysis of symptoms only is likely to be poorly informative given the very large heterogeneity of major depression clinical profile (26). An alternative approach is to identify biologically-driven subtypes by grouping patients based on shared neuro-biological features (27). Recent studies applied clustering techniques on resting-state (28-31) and structural neuroimaging data (32, 33), providing new insights in how distinct neurobiological signatures could be linked to various clinical manifestations of depression. Despite the variability in the number of subtypes, most of the studies converge on the identification of anxiety-, anhedonia-, and insomnia-related data-driven clusters, as well as on clusters’ differences in depression severity and recurrence (34). Some of these “biotypes” were found to be also predictive of response to antidepressant treatments (28), but they failed to be replicated in independent samples (35). Indeed, a common difficulty in the implementation of unsupervised algorithms is the lack of ground truth information and the absence of a defined approach to assess clustering validity, which further complicates the replicability of subtypes nested within clinical populations (36, 37). Moreover, none of these studies combined different neuroimaging modalities to investigate whether multimodal information could enhance the discovery of relevant subgroups.

In the current study, we aim to discover biologically-driven subtypes of patients with different levels of antidepressant treatment response based on multimodal structural neuroimaging, including grey matter volumes, cortical thickness, and extracted fractional anisotropy (FA) values of white matter tracts, in MDD patients. We implemented a novel analytical pipeline including data harmonization, dimensionality reduction, and stability-based relative clustering validation to identify clusters of patients that are stable and generalizable.
to unseen observations (38). The identified data-driven clusters were then profiled for clinical variables of interest, including TRD, depressive symptomatology, and history of childhood trauma.

Materials and Methods

Participants

A total of 102 patients with a current diagnosis of MDD and an ongoing depressive episode were recruited at IRCCS San Raffaele Scientific Institute (Milan, Italy). Detailed inclusion and exclusion criteria are reported in Methods S1. After a complete description of the study, written informed consent was obtained. All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The study is approved by the local ethical committee.

Clinical and sociodemographic measures

Age, sex, number of previous mood episodes, age of onset, duration of illness, body mass index (BMI), and years of education were collected. To obtain a measure for pharmacological treatment, medication load was calculated for each subject categorizing each medication into low-dose to high-dose groupings, scored as 0 (no medication), 1–4 (low to high dosages) as described by Sackeim (39) (see Methods S2 for detailed criteria). Treatment resistance was defined as a failure to respond to at least two antidepressant treatments of adequate dosages and duration (3, 7). Clinical data were assessed by the psychiatrist in charge using best estimation procedure, taking into account available charts, case notes, and information provided by at least one relative (40). In a subsample of 64 patients, severity of depression was rated on the Beck Depression Inventory–Short Form (BDI-SF) (41), and composite scores for different domains (Negative Self-Esteem, Anergy, and Dysphoria) were derived (42). Childhood trauma was evaluated using the 28-items Childhood Trauma Questionnaire (CTQ) (43). CTQ is a self-administrated inventory that was developed for a reliable retrospective assessment of neglect and abuse during childhood. The questionnaire is composed of 5 subscales that evaluate different aspects of neglect and abuse (physical, sexual, and emotional abuse, and physical and emotional neglect). Moreover, a Minimization/Denial validity scale was used to identify underreported
maltreatment (44).

**MRI data acquisition and pre-processing**

All subjects underwent a magnetic resonance scan at C.E.R.M.A.C. (Centro di Eccellenza Risonanza Magnetica ad Alto Campo, University Vita-Salute San Raffaele, Milan, Italy). T1-weighted and diffusion tensor images (DTI) were acquired on two 3.0 Tesla scanners. Technical details for the two scanners are reported in Methods S3.

T1-weighted neuroanatomical images were processed using the Computational Anatomy Toolbox (CAT12) for SPM (45). T1 images were normalized to an anatomical model and segmented into gray matter, white matter, and cerebrospinal fluid (CSF). Check of spatial alignment and sample homogeneity was performed to exclude outliers. Then, low-pass spatial filtering (smoothing) techniques were used to remove any potential artefacts in the tissue maps. Cortical thickness was derived for 68 regions defined by the Desikan-Killiany atlas (46) and 148 from the Destrieux atlas (47), whereas volumes of cortical and subcortical grey matter volumes were extracted for 122 regions labelled by the Neuromorphometrics atlas (http://Neuromorphometrics.com/). Finally, total intracranial volume (TIV) was computed. Whole-brain tract-wise average fractional anisotropy (FA) values were extracted according to ENIGMA-DTI protocols (http://enigma.ini.usc.edu/protocols/dtiprotocols/). DTI images were pre-processed using FMRIB Software Library (FSL) tools. Specifically, all volumes were corrected for eddy current induced distortions and subjects’ movements (48). Then, a brain mask was created using Brain Extraction Tool (BET) (49), which deletes non-brain tissues from the image. Next, by FSL’s DTIFIT command, included in FDT (50), a voxel-wise diffusion tensor model was fit to the data in order to obtain parametric maps of FA. All subject’s FA images were then processed using FSL’s TBSS analytic method and were aligned to the MNI space, by using local deformation procedures performed by FMRIB’s FNIRT. The mean of all aligned FA was then created, and a “thinning” process was applied to create a skeletonized mean FA image representing the centers of all common tracts. A threshold of 0.2 was set to this image in order to control for intersubject variability and reduce the likelihood of partial volume effect. Quality control, including inspections of data, vector gradients, registration, and average skeleton projection distance, were performed according to the ENIGMA-DTI protocol. Finally, all individual FA images were projected onto the skeleton by searching perpendicular from
the skeleton for maximum FA values (51). Average FA values were calculated from voxels in each subject’s white matter skeleton within 63 tract-wise ROIs, derived from the Johns Hopkins University (JHU) white matter parcellation atlas (52). FA values were separately calculated for right and left hemispheres in each ROI, except for body of corpus callosum (BCC), corpus callosum (CC), fornix (FX), genu of corpus callosum (GCC) and splenium of corpus callosum (SCC).

Since MRI data were acquired with two different scanners, we implemented the ComBat algorithm to correct for potential technical artefacts (i.e., “batch effects”) that could affect meaningful relationships between biological signal and clinical outcomes, leading to unreliable conclusions (53). The ComBat algorithm was adapted to be performed on regional gray matter estimates and extracted tract-based FA values. Of note, age, sex, and TIV (only for grey matter measures) were considered as biological covariates in order to be protected from the removal of scanner effects.

Statistical analysis

Stability-based relative clustering validation

Clustering analyses were performed using the reval Python library (https://github.com/IIT-LAND/reval_clustering), which implements a stability-based relative clustering approach within a cross-validation framework to identify the clustering solution that best replicates on unseen data. (38) (For further details, see Methods S4). For the current study, reval was adapted to be used on neuroimaging features, choosing a Gaussian mixture model (GMM) with full covariance matrix as clustering algorithm, and support vector machine (SVM) as a supervised classifier. Given the high-dimensionality of the data, we performed the analyses both with and without dimensionality reduction to investigate the impact of multidimensionality on clustering performance (github repository: https://github.com/fede-colombo/NeuReval).

In the first analysis, data were mean centered and normalized to the same scale. Input variables were adjusted for confounding effects of age, sex and TIV (only in grey matter) using linear regression. All these steps were fitted on the training set and then applied to the test set. A 2-folds cross-validation scheme (50% training, 50% test) was implemented to control for the size imbalance that derives from training-test splitting. The cross-validation scheme was repeated 10 times with 10 random labelling iterations to ensure robustness. To optimize the SVM hyperparameters (i.e., soft-margin C ranged 0.01, 0.1, 1, 10, 100, 1000) and the type of kernel (i.e.,
linear or radial), a grid search was performed in the training set and selection was based on normalized stability. Since previous studies identified from 2 to 4 clusters of MDD based on neuroimaging data (34), the entire cross-validation procedure was iterated over a number of GMM components from 2 to 5. The best clustering solution was the one that minimized the normalized stability throughout cross-validation. To assess the internal validity of the identified clusters, we also computed the silhouette and Davies-Bouldin scores by applying GMM on the entire dataset and with the same range of number of components used in the stability-based regime, and the adjusted mutual information (AMI) index was calculated to evaluate the similarity of clusters’ labels derived from the stability-based approach and the ones obtained from silhouette and Davies-Bouldin scores model’s selection.

The same analysis was repeated by applying Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction. UMAP was initialised with a number of neighbours equal to 30, minimum distance equal to 0.0, and Euclidean distance metric, as suggested in the documentation (https://umap-learn.readthedocs.io/en/latest/clustering.html). Together with SVM hyperparameters, also the number of UMAP components was optimized within cross-validation through grid search in the training set, iterating over 2 to 5 components.

Despite the stability-based relative clustering approach allows us to assess the stability of a given clustering solution by means of cross-validation, it does not assess whether the identified clusters reflect the existence of true subgroups within the data. Therefore, we further investigated whether the selected clustering solutions significantly deviated from the null hypothesis that data derive from a single multivariate Gaussian distribution computed from 10,000 Monte Carlo simulations using the sigclust library in R (54).

Clinical, demographic, and neuroimaging characterization of the clusters

For both the analyses, the identified clusters were compared for neuroimaging features, TRD, age, sex, education, type of scanner, number of episodes, age of onset, duration of illness, medication load, BMI, CTQ total score and subscales, BDI total scores and domains, and HDRS-21 total score. Two-sample t-tests were performed for continuous variables, whereas $\chi^2$ was calculated for categorical variables. For neuroimaging data, Cohen’s d effects sizes were also computed and comparisons were deemed as significant if they passed a false-discovery rate (FDR) $q<0.05$ threshold to correct for multiple comparisons.
To investigate whether the data-driven clusters reflect clinically-meaningful subgroups of depressed patients, we performed multivariate analysis of variance (MANOVA) considering BDI-SF domains, and CTQ subscales as dependent variables and clusters’ labels as fixed factors. Age, sex, number of previous episodes, and medication load were entered as covariates to control for potential confounding effects. In case of significant results, subsequent linear discriminant analysis (LDA) was performed to assess how the dependent variables discriminate the clusters. All the analyses were performed in R using the mancova (jmv library) and lda (MASS library) functions.

Results

Unsupervised data-driven identification of depression clusters based on neuroimaging data

The clinical and demographic characteristics of the sample are shown in Table 1. The stability-based relative validation clustering approach on multimodal structural neuroimaging data without data reduction identified that a 2-clusters solution minimized the normalized stability (0.316), showing a good discriminative performance with an accuracy of 68.4% (Figure 2b) (Table 2). The identified clustering solution deviated from the null hypothesis that clusters were generated from one single Gaussian distribution (p=0.035). Density plot of the log probabilities estimated by the GMM model showed that, although the subtypes distributions are partially overlapping, they reflected two distinct Gaussian distributions (Figure 2A). Similar results were obtained with UMAP dimensionality reduction (Table 2, Figure S1A and S1B).

Neurobiological-driven clusters are related to treatment response and specific neurobiological profiles

The identified clusters without UMAP data reduction were significantly different for TRD (p=0.008), sex (p=0.03), and education (p=0.016). Cluster 1 (N=43) was mainly composed by treatment-responsive patients whereas Cluster 2 (N=59) was characterised by higher TRD rates (Figure 2d). Notably, a higher F:M ratio was observed in Cluster 2 compared to Cluster 1. No significant differences in age, number of episodes, age of onset, duration of illness, and BMI was found between the two clusters (Table 3). Cortical thickness in temporo-parietal structures was significantly higher in Cluster 1 compared to Cluster 2 with large effect (pFDR < 0.001) (Figure 3A). Considering gray matter volumes, the clusters were different with large effects in the bilateral thalami and parahippocampal gyrus, right middle occipital gyrus, right superior occipital gyrus, left
temporal pole, right precuneus, right postcentral gyrus, left medial frontal cortex, left gyrus rectus, left precentral gyrus, and left superior parietal lobule, with higher volumes in Cluster 1 compared to Cluster 2 ($p_{\text{FDR}} < 0.001$) (Figure 3b). As for white matter, higher FA values were found in Cluster 1 compared to Cluster 2 (Figure 3c). Specifically, the right superior fronto-occipital fasciculus ($p_{\text{FDR}} = 0.032$), the body of the corpus callosum ($p_{\text{FDR}} = 0.024$), and the right stria terminalis ($p_{\text{FDR}} = 0.037$) differentiated the two clusters with a moderate effect. (Table S1). A similar clustering solution was identified in the UMAP-reduced model (Results S1).

Neuroimaging-based subtypes reflect specific childhood trauma and depression-related signatures

The identified data-driven clusters reflected subtypes of depressed patients based on childhood trauma and depressive symptomatology using a multivariate approach. MANOVA results showed a significant between-clusters difference (Wilks’ lambda=0.68, $F(9,50)=2.64$, $p=0.014$), considering age, sex, medication load, and number of episodes as nuisance covariates. Results from univariate statistics were not significant (Table 3). The following LDA identified one discriminant function (variate) that significantly differentiated the two clusters (Wilks’ lambda=0.70, $p=0.013$). Standardised discriminant coefficients revealed that BDI Anergy ($b=1.48$), CTQ minimisation/denial ($b=1.23$), and CTQ emotional neglect ($b=1.07$) were largely positively associated with the variate, whereas Dysphoria ($b=-0.65$), Negative Self-Esteem ($b=-0.59$), and CTQ physical neglect ($b=-0.36$) showed a negative relationship with the variate. These results suggest that higher scores on the BDI Anergy, CTQ minimisation/denial and emotional neglect domains indicate a higher probability to be assigned to Cluster 2. Conversely, patients with higher scores on the BDI Dysphoria, Negative Self-Esteem, and CTQ physical neglect are more likely to belong to Cluster 1 (Figure 2c). Similar findings were also obtained using clusters’ labels derived from the UMAP-reduced model (Results S1).

Discussion

By implementing a novel cross-validated clustering procedure, we identified two subgroups of patients driven from grey matter and DTI that were differentially associated with treatment resistance, depressive symptomatology and childhood trauma. In particular, one cluster showed a mixed profile with a higher proportion of treatment-resistant patients (high-TRD), higher F:M sex ratio, and more likely to be associated
with energy-related depressive symptomatology and history of childhood emotional neglect and abuse. A second cluster was largely characterised by treatment-responsive patients (low-TRD), with higher cognitive and affective symptoms of depression. In terms of neuroimaging profiles, the high-TRD cluster displayed reduced widespread cortical thickness and volumes compared to the low-TRD one, along with lower white matter integrity. The identified clusters were differentiable with an accuracy of 67%, suggesting that structural neuroimaging features can enhance the discovery of clinically-meaningful subtypes of depression and treatment resistant patients.

Considering the clinical signatures of the identified subgroups, the BDI Anergy domain contributed the most in differentiating the two groups, suggesting a positive relationship with the high-TRD cluster and symptoms related to fatigue, decreased appetite, and work difficulties, especially in females. Previous studies showed that symptoms reflecting an altered energy intake/output balance (e.g., appetite changes, sleep disturbances, and fatigue), characterize a specific MDD subtype characterised by altered immune-metabolic functions (55, 56). In line with our stratification model, convergent evidence indicates that this symptomatology is more frequent in females with an early onset of diseases, recurrent episodes, and poor treatment outcomes (57). Also CTQ minimization/denial, emotional abuse and neglect scores contributed to discrimination between the two subtypes, reflecting a higher probability to belong to the high-TRD subtype. Previous evidence showed a possible link between childhood trauma and neurovegetative depressive symptomatology, reporting higher exposure to traumatic events in MDD patients with atypical features compared to those without (58). Indeed, it is possible that the high-TRD cluster identifies a specific MDD subgroup with a higher sensitivity to early stress and more current anergic symptoms, in analogy to the extensive research documenting higher prevalence of anergic/atypical symptoms in reactive depression, both categories being less responsive to monoamine reuptake inhibitors (59, 60). This symptom profile and childhood trauma may also suggest a subthreshold bipolarity, usually associated with TRD (61).

The data-driven subgroups also displayed distinct grey matter and structural connectivity profiles. Volumetric reductions in the precuneus, parahippocampal gyrus, thalamus, temporal pole, and pre- and post-central gyri, were found in the high-TRD subtype. Most of these structures represent core areas in cortico-striatal-thalamic circuits, which are involved in attribution of salience to external stimuli, emotional regulation of inner states, and cognition (62, 63). Aberrant function and connectivity within these networks has been previously
associated with treatment resistance (64, 65). Notably, a similar stratification was found based on functional neuroimaging features, with hyperconnectivity in thalamic and fronto-striatal networks associated with increased anhedonia and psychomotor retardation and less responsiveness to TMS intervention (28). Our study provides additional evidence supporting these findings, indicating that these functional abnormalities might reflect morphological alterations. When considering cortical thickness, the largest effects sizes were observed for temporo-parietal structures, with higher thickness values in the low-TRD subtype. Similar results were observed in a larger cohort study, where two MDD subgroups based on cortical thickness and surface reflected differences on general cognitive ability (33). Our results take a step forward providing evidence of an association with treatment response and other clinically meaningful measures. Together with grey matter, lower FA values in the superior fronto-occipital fasciculus, body of corpus callosum, stria terminalis, and hippocampal cingulum, were found in the high-TRD subtype. Previous studies brought evidence of decreased FA affecting these structures in TRD patients compared to healthy controls and first-episode MDD patients (66, 67), whereas microstructural changes in the cingulum and stria terminalis are predictive of remission after 8-weeks of antidepressant treatment (68). Accordingly, the preserved integrity of these structures might reflect a better cortico-limbic modulation and inter-hemispheric integration, enhancing probability of remission after treatment.

Previous studies aimed at identifying subgroups of MDD patients using neuroimaging data, showing how the clinical heterogeneity of MDD can be disentangled based on shared neurobiological correlates (27-29, 69). By exploring the relationships with childhood trauma and treatment response, we demonstrated that the identified biologically-driven subtypes are informative about relevant clinical features outside depressive symptomatology, providing a more comprehensive picture of the multifaceted manifestations of MDD. A large drawback of previous studies is the application of conventional clustering approaches tied to the specific data at hand, limiting the generalizability of the identified subtypes. Contrary, the clustering pipeline implemented in our study has the advantage to translate the unsupervised setting into a supervised classification problem, providing a measure of clusters’ stability and replicability to unseen data through cross-validation. Furthermore, a high-value strength of this work is the clinical validation of the discovered clusters, which can potentially provide insights for new effective treatments tailored on the patient’s biological and clinical profile. For instance, patients assigned to the high-TRD cluster might benefit from antidepressant therapies inducing...
neuroplastic changes, such as electroconvulsive therapy (70-73), repetitive transcranial magnetic stimulation (74, 75), deep brain stimulation (76, 77), and ketamine (78-80), whereas treatments targeting the immune-metabolic pathways, such as physical exercise, diet or anti-inflammatory treatments, may be also useful in alleviating the energy-related symptoms (81-83). Furthermore, considering that the high-TRD cluster was also associated with high scores for CTQ, this group may benefit from adjunctive psychotherapeutic interventions focused on trauma (84, 85). On the other hand, psychotherapies targeting cognitive distortions, such as cognitive behavioural therapy, might represent optimal treatment options for patients assigned to the low-TRD group (86-88).

The limitations of the current study should be acknowledged. First, all patients were recruited in the same clinical center, limiting the possibility to extend our findings to other cohorts. However, the cross-validation framework provides a good approximation of clusters’ generalizability (38). Second, since all patients were under pharmacological treatment at the time of scanning, we cannot rule out long-lasting effects of drugs on neuroimaging features. Another limitation comes from the sample size, which further may limit results generalizability. Our stratification model should be evaluated in larger cohorts to provide realistic applications in clinical practice. Finally, we recognize that noise in biological data might have affected clustering solutions. It should be considered, though, that the pre-processing steps employed (e.g., ComBat harmonization) helped in mitigating this issue.

In conclusion, our results indicate that structural neuroimaging data can be used to define novel subtypes of depression that are informative about treatment resistance, depressive symptomatology, and childhood trauma. A multimodal stratification including other kind of data (e.g., functional neuroimaging, genetic data, and immune-inflammatory markers) may improve the prediction of disease’s outcomes, uncovering how different pathophysiological mechanisms interact with different clinical features of depression.

Acknowledgments

This study was supported by the Italian Ministry of Health (grant numbers GR 2019-12370616 and PNRR-MAD-2022-12375859).
Disclosures

The authors declare no financial interests or potential conflicts of interest. CF was a speaker for Janssen. AS is or was a consultant/speaker for Abbott, Abbvie, Angelini, AstraZeneca, Clinical Data, Boehringer, Bristol-Myers Squibb, Eli Lilly, GlaxoSmithKline, Innovapharma, Italfarmaco, Janssen, Lundbeck, Naurex, Pfizer, Polifarma, Sanofi, Taliaz and Servier.
References


labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage. 31:968-980.


### Table 1. Descriptive statistics of the whole sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>49.43 (10.00)</td>
</tr>
<tr>
<td>Sex</td>
<td>33 M, 69 F</td>
</tr>
<tr>
<td>scanner</td>
<td>46 old, 56 new*</td>
</tr>
<tr>
<td>number of episodes</td>
<td>5.67 (6.21)</td>
</tr>
<tr>
<td>age of onset</td>
<td>32.30 (11.78)</td>
</tr>
<tr>
<td>duration of illness</td>
<td>17.13 (10.51)</td>
</tr>
<tr>
<td>Education years</td>
<td>12.75 (3.82)</td>
</tr>
<tr>
<td>medication load</td>
<td>4.57 (2.03)</td>
</tr>
<tr>
<td>BMI</td>
<td>24.82 (5.27)</td>
</tr>
<tr>
<td>TRD</td>
<td>71 non-TRD, 31 TRD</td>
</tr>
<tr>
<td>HDRS-21 total score</td>
<td>21.03 (5.91)</td>
</tr>
</tbody>
</table>

Abbreviations: SD, standard deviation; BMI, body mass index; TRD, treatment-resistant depression; HDRS, Hamilton Depression Rating Scale. * T1-weighted and diffusion tensor images were acquired on two 3.0 Tesla scanners. Until June 2016, 46 MDD patients were acquired with a Gyroscan Intera, Philips (old scanner). From October 2016, 56 MDD patients were examined with the Ingenia CX, Philips (new scanner).

### Table 2. Performances and hyper-parameters for the best clustering solutions

<table>
<thead>
<tr>
<th>Model</th>
<th>Clusters</th>
<th>UMAP components</th>
<th>SVM hyperparameters</th>
<th>Normalized stability (error)</th>
<th>Accuracy</th>
<th>Silhouette score (AMI)</th>
<th>Davies-Bouldin score (AMI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM + SVM</td>
<td>2</td>
<td>NA</td>
<td>C=1, kernel=rbf</td>
<td>0.316 (0.316, 0.098)</td>
<td>0.684</td>
<td>0.167 (AMI=0.482)</td>
<td>1.89 (AMI=0.482)</td>
</tr>
<tr>
<td>UMAP + GMM + SVM</td>
<td>2</td>
<td>4</td>
<td>C=0.1, kernel=linear</td>
<td>0.333 (0.333, 0.067)</td>
<td>0.667</td>
<td>0.508 (AMI=0.509)</td>
<td>0.626 (AMI=0.509)</td>
</tr>
</tbody>
</table>

Abbreviations: GMM, Gaussian Mixture Model; SVM, Support Vector Machine; UMAP, Uniform Manifold Approximation and Projection; AMI, adjusted mutual information index.
Table 3. Demographic and clinical characteristics of the identified data-driven clusters without UMAP dimensionality reduction.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cluster 1 (N=43)</th>
<th>Cluster 2 (N=59)</th>
<th>t / χ²</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>50.98 (8.73)</td>
<td>48.31 (10.77)</td>
<td>1.34</td>
<td>0.184</td>
</tr>
<tr>
<td>sex</td>
<td>19 M, 24 F</td>
<td>14 M, 45 F</td>
<td>4.76</td>
<td>0.03*</td>
</tr>
<tr>
<td>scanner</td>
<td>19 old, 24 new</td>
<td>27 old, 32 new</td>
<td>0.03</td>
<td>0.874</td>
</tr>
<tr>
<td>number of episodes</td>
<td>5.47 (5.15)</td>
<td>5.81 (6.91)</td>
<td>-0.28</td>
<td>0.781</td>
</tr>
<tr>
<td>age of onset</td>
<td>33.93 (11.18)</td>
<td>31.12 (12.15)</td>
<td>1.19</td>
<td>0.236</td>
</tr>
<tr>
<td>duration of illness</td>
<td>17.05 (11.05)</td>
<td>17.19 (10.19)</td>
<td>-0.07</td>
<td>0.947</td>
</tr>
<tr>
<td>Education years</td>
<td>11.70 (3.80)</td>
<td>13.53 (3.68)</td>
<td>-2.44</td>
<td>0.016*</td>
</tr>
<tr>
<td>medication load</td>
<td>4.35 (2.13)</td>
<td>4.71 (2.14)</td>
<td>-0.85</td>
<td>0.398</td>
</tr>
<tr>
<td>BMI</td>
<td>25.29 (4.63)</td>
<td>24.99 (4.76)</td>
<td>0.32</td>
<td>0.375</td>
</tr>
<tr>
<td>TRD</td>
<td>36 non-TRD, 7 TRD</td>
<td>35 non-TRD, 24 TRD</td>
<td>7.00</td>
<td>0.008**</td>
</tr>
<tr>
<td>HDRS-21 total score</td>
<td>23.19 (7.26)</td>
<td>22.27 (7.00)</td>
<td>0.64</td>
<td>0.522</td>
</tr>
<tr>
<td>CTQ total score</td>
<td>44.92 (15.04)</td>
<td>41.40 (13.41)</td>
<td>0.94</td>
<td>0.351</td>
</tr>
<tr>
<td>CTQ physical abuse</td>
<td>6.25 (2.47)</td>
<td>6.13 (2.46)</td>
<td>0.20</td>
<td>0.845</td>
</tr>
<tr>
<td>CTQ emotional neglect</td>
<td>13.83 (5.54)</td>
<td>12.90 (5.27)</td>
<td>0.67</td>
<td>0.504</td>
</tr>
<tr>
<td>CTQ emotional abuse</td>
<td>9.38 (4.54)</td>
<td>8.93 (4.31)</td>
<td>0.40</td>
<td>0.693</td>
</tr>
<tr>
<td>CTQ physical neglect</td>
<td>8.54 (4.03)</td>
<td>7.03 (3.04)</td>
<td>1.71</td>
<td>0.093</td>
</tr>
<tr>
<td>CTQ sexual abuse</td>
<td>6.08 (3.51)</td>
<td>6.18 (3.38)</td>
<td>-0.10</td>
<td>0.918</td>
</tr>
<tr>
<td>CTQ minimization/denial</td>
<td>7.92 (3.13)</td>
<td>9.40 (3.05)</td>
<td>-1.86</td>
<td>0.067</td>
</tr>
<tr>
<td>BDI total score</td>
<td>16.13 (8.96)</td>
<td>16.73 (7.74)</td>
<td>-0.28</td>
<td>0.778</td>
</tr>
<tr>
<td>BDI Self-Esteem</td>
<td>5.92 (3.48)</td>
<td>5.80 (3.08)</td>
<td>0.14</td>
<td>0.889</td>
</tr>
<tr>
<td>BDI Anergy</td>
<td>4.79 (3.02)</td>
<td>6.15 (2.53)</td>
<td>-1.93</td>
<td>0.058</td>
</tr>
<tr>
<td>BDI Dysphoria</td>
<td>5.42 (3.62)</td>
<td>5.20 (2.70)</td>
<td>0.27</td>
<td>0.784</td>
</tr>
</tbody>
</table>

Results for continuous variables are reported as mean (standard deviation), whereas sample size is reported for categorical variables. Abbreviations: BMI, body mass index; TRD, treatment-resistant depression; HDRS-21, 21-Hamilton Depression Rating Scale; CTQ, Childhood Trauma Questionnaire; BDI, Beck Depression Inventory.

*a subsample of 64 MDD patients with clinical scales (cluster 1: N=24; cluster 2: N=40). *p<0.05, **p<0.001
Figure 1. Schematic representation of the statistical analysis plan. A) reval algorithm pipeline. B) analysis pipeline for applying reval on neuroimaging data. Clusters’ labels derived from the best clustering solution (i.e., minimum normalized stability) were used for MANOVA and clusters’ comparisons on neuroimaging features and treatment response. Adapted from (38).
**Figure 2. Unsupervised data-driven identification of depression subtypes based on neuroimaging data.**

This figure represents results from the stability-based relative clustering approach. A) Probability density distributions identified by Gaussian Mixture Model and Support Vector Machine without UMAP data reduction. B) Normalized stability plot. Error bars represent the 95% confidence interval for normalized stability from the 10x2 repeated cross-validation. The optimal number of clusters that minimizes the normalized stability was 2, achieving an accuracy of 68%. C) Standardized coefficients of the linear discriminant analysis considering clusters’ labels as fixed factors and BDI and CTQ domains as dependent variables. Positive weights (orange) reflect a higher probability to be assigned to the high-TRD cluster, whereas negative weights (green) indicate a higher probability to belong to the low-TRD cluster. D) Graphical representation of the proportion of TRD patients between the identified clusters.
Figure 3. Standardized effect sizes differences (Cohen’s d) in cortical, subcortical, and white matter structures associated with the identified clusters. Increasing red color indicates higher effect sizes (low-TRD>high-TRD). A) Effect sizes for cortical thickness based on Destrieux (top) and Desikan-Killiany (down) atlases. B) Effect sizes for extracted tract-based FA values. C) Effect sizes for extracted grey matter volumes.