Title

Using structural MRI to identify bipolar disorders — 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group

Journal

Molecular Psychiatry, 25(9)

ISSN

1359-4184

Authors

Nunes, Abraham

Schnack, Hugo G

Ching, Christopher RK

et al.

Publication Date

2020-09-01

DOI

10.1038/s41380-018-0228-9

Molecular Psychiatry (2020) 25:2130–2143
https://doi.org/10.1038/s41380-018-0228-9

ARTICLE

Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group

Abraham Nunes[1,2] et al. · for the ENIGMA Bipolar Disorders Working Group

Received: 15 February 2018 / Revised: 11 June 2018 / Accepted: 24 July 2018 / Published online: 31 August 2018

Abstract

Bipolar disorders (BDs) are among the leading causes of morbidity and disability. Objective biological markers, such as those based on brain imaging, could aid in clinical management of BD. Machine learning (ML) brings neuroimaging analyses to individual subject level and may potentially allow for their diagnostic use. However, fair and optimal application of ML requires large, multi-site datasets. We applied ML (support vector machines) to MRI data (regional cortical thickness, surface area, subcortical volumes) from 853 BD and 2167 control participants from 13 cohorts in the ENIGMA consortium. We attempted to differentiate BD from control participants, investigated different data handling strategies and studied the neuroimaging/clinical features most important for classification. Individual site accuracies ranged from 45.23% to 81.07%. Aggregate subject-level analyses yielded the highest accuracy (65.23%, 95% CI = 63.47–67.00, ROC-AUC = 71.49%, 95% CI = 69.39–73.59), followed by leave-one-site-out cross-validation (accuracy = 58.67%, 95% CI = 56.70–60.63). Meta-analysis of individual site accuracies did not provide above chance results. There was substantial agreement between the regions that contributed to identification of BD participants in the best performing site and in the aggregate dataset (Cohen's Kappa = 0.83, 95% CI = 0.829–0.831). Treatment with anticonvulsants and age were associated with greater odds of correct classification. Although short of the 80% clinically relevant accuracy threshold, the results are promising and provide a fair and realistic estimate of classification performance, which can be achieved in a large, ecologically valid, multi-site sample of BD participants based on regional neurostructural measures. Furthermore, the significant classification in different samples was based on plausible and similar neuroanatomical features. Future multi-site studies should move towards sharing of raw/voxelwise neuroimaging data.

Introduction

Bipolar disorders (BDs) are lifelong conditions, which tend to start in adolescence or early adulthood and consequently rank among the leading causes of morbidity and disability worldwide [1, 2]. Despite substantial advances in our understanding of the neurobiology of BD, the diagnostic system in psychiatry continues to be based on description of behavioral symptoms. This often results in delayed or inaccurate diagnosis of BD [3–5], which in turn leads to delayed or ineffective treatment [6]. Objective, biological markers could aid significantly in the clinical management of mental disorders [7], might reduce stigma, facilitate research and expedite the development of new treatments [8].

Brain imaging offers the unique ability to non-invasively investigate brain structure and function. Previous brain-imaging meta-analyses and large-scale multi-site studies have demonstrated that adults with BD had robust and replicable neurostructural alterations in subcortical, that is, hippocampus, amygdala, thalamus [9–11], as well as cortical regions, including inferior frontal gyrus, precentral gyrus, fusiform gyrus, middle frontal cortex [12–14]. Despite these advances and the relatively broad availability, the diagnostic potential of magnetic resonance imaging (MRI) in psychiatry has not been fully realized.

The translation of brain imaging from bench to the bedside has been hindered by the low sensitivity and

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41380-018-0228-9) contains supplementary material, which is available to authorized users.

✉ Tomas Hajek
tomas.hajek@dal.ca

Extended author information available on the last page of the article

specificity of between-group differences, by clinical heterogeneity and limited generalizability of findings from relatively small samples. The problem of low sensitivity and specificity may be overcome by novel analytical tools, such as machine learning (ML) [15, 16]. Traditional mass-univariate methods of MRI data analysis focus on localized and spatially segregated patterns of between-group differences [17]. The effect sizes of such changes (Cohen's $d =$ 0.15–0.29 [11, 14]) tend to be many times smaller than the effects needed for clinical application (Cohen's $d = 1.50$–3.00 [18, 19]). In contrast, the ML techniques increase predictive power by targeting multivariate alterations distributed throughout the whole brain, which may better characterize the abnormalities found in psychiatric disorders [15, 20, 21]. Thus, ML brings neuroimaging analyses to the level of individual subjects, and with some caveats, potentially allows for diagnostic use. When previously applied to structural MRI, ML differentiated BD from control participants with accuracies between 59.5% [22] and 73.00% [23].

However, ML approaches typically require large samples to optimize the performance of the classifier, provide a generalizable snapshot of the studied disorder, decrease the risk of sampling effects and allow for application of rigorous cross-validation approaches [19]. Single-site studies may provide high site-specific accuracies [24], which, however, may not generalize across samples [25, 26]. Small studies may also yield a wide range of classification performances and inconsistencies in regions, which contribute to the overall classification [25–27]. Previous ML structural MRI studies in BD have typically included <50 BD participants recruited in a single site [23, 28–34]. The largest currently available neurostructural ML studies investigated 128–190 BD and 127–284 control participants [35–37], from up to two sites [22, 23, 38].

Large, multi-site datasets will necessarily be more heterogeneous than single site, carefully controlled samples. In fact, heterogeneity is one of the defining characteristics of big-data [39]. Single-site studies with rigorous inclusion/exclusion criteria may help isolate sources of heterogeneity, but they will represent only a small fraction of the "patient space." In contrast, a large, multi-site study will primarily target generalizable alterations, which are shared among the participants, regardless of illness subtype, effects of treatment and other clinical variables. This is related to the fact that different sources of heterogeneity (i.e., presence of psychosis, neuroprogression, exposure to medications) affect different brain regions and often act in opposing directions [11–14, 40–42]. In addition, individual sources of heterogeneity, which are present only in some participants, are unlikely to systematically bias the findings in large, multi-site investigations. Thus, smaller, carefully controlled studies and large, multi-site datasets are complementary and ask different questions. BD is a broad and heterogeneous condition. Therefore, it is all the more important to quantify the extent to which ML can classify large, ecologically valid datasets based on neuroanatomy.

Researching generalizable brain alterations has only recently become possible through research consortia committed to aggregation and sharing of brain-imaging data across research groups. Despite the inherent limitations, retrospective data sharing initiatives create an optimal environment for application of ML strategies and for a fair and realistic estimation of the utility of MRI for classification of neuropsychiatric disorders. This approach has been utilized to improve predictive models of autism or Alzheimer dementia [26], but has not yet been applied to BD. The Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) consortium is an international multi-cohort collaboration, which, by combining datasets from multiple sites, has allowed for more accurate testing of the reproducibility of disease effects in participants with schizophrenia [43], BD [11, 14] or major depression [44]. Due to the multi-site nature, methodological harmonization and access to some of the largest neuroimaging datasets to date, the ENIGMA platform provides an ideal opportunity to test ML on sufficiently large and generalizable samples.

In collaboration with the ENIGMA-BD Working Group, we applied ML to structural MRI data from 3020 participants recruited in 13 independent sites around the world. We attempted to differentiate BD from control participants based on brain structure. In addition, we studied the effects of different data handling strategies on classification performance, described the neuroanatomical features, which contributed to individual subject classification and investigated the effects of clinical variables on classification performance.

Materials and methods

Samples

The ENIGMA-BD Working Group brings together researchers with brain imaging and clinical data from BD participants and healthy controls [11, 14]. Thirteen of the sites from previously published ENIGMA studies [11, 14] provided individual subject data for this ML project. Each cohort's demographics are detailed in Supplementary Table S1. Supplementary Table S2 lists the instruments used to obtain diagnosis and clinical information. Supplementary Table S3 lists exclusion criteria for study enrollment. Briefly, all studies used standard diagnostic instruments, including SCID ($N = 10$), MINI ($N = 2$) and DIGS ($N = 1$). Most studies ($N = 7$) included bipolar spectrum disorders, five studies included only BD-I and a

single study only BD-II participants. Substance abuse was an exclusion criterion in 9/13 studies. Most studies (10/13) did not exclude comorbidities, other than substance abuse. A single study recruited medication naive participants. The remaining studies did not restrict medication use. Consequently, the sample is a broad, ecologically valid and generalizable representation of BD.

All participating sites obtained approval from their local institutional review boards and ethics committees, and all study participants provided written informed consent.

Image processing and analyses

Structural T1-weighted MRI brain scans were acquired at each site and analyzed locally using harmonized analysis and quality control protocols from the ENIGMA consortium. Image acquisition parameters are listed in Supplementary Table S4. All groups used the same analytical protocol and performed the same visual and statistical quality assessment. These harmonized protocols were used in the previous publications by our group [11, 14] and they have been applied more broadly in large-scale ENIGMA studies of other disorders. Briefly, using a freely available and extensively validated FreeSurfer software, we performed segmentations and parcellations into 7 subcortical and 34 cortical gray matter regions per hemisphere (left and right), based on the Desikan–Killiany atlas. Visual quality controls were performed on a region of interest (ROI) level aided by a visual inspection guide including pass/fail segmentation examples. Diagnostic histogram plots were generated for each site and outlier subjects were flagged for further review. All ROIs failing quality inspection were withheld from subsequent analyses. Previous analyses from the ENIGMA-BD Working Group showed that scanner field strength, voxel volume and the version of FreeSurfer used for segmentation did not significantly influence the effect size estimates. Further details regarding these analyses, as well as forest plots of cortical and subcortical effect sizes from individual sites can be found here [11, 14].

Data preprocessing

Input features were ROI cortical thicknesses (CT), surface area (SA) and subcortical volumes, a total of 150 features, and intracranial volume. As SA and CT are genetically distinct [45], influenced by different neurobiological mechanisms [46] and sometimes affected in opposite directions [47], we used both as input features. Prior to fitting of the ML classifier, we imputed missing data using mean values of the respective features, as well as centered and scaled each continuous feature.

Using statistical harmonization to reduce heterogeneity of data could improve accuracy [48], but at a cost to generalizability. Such approaches may compromise the train/test separation and may introduce additional assumptions, which are difficult to verify. Thus, in keeping with other studies [23, 38, 49], instead of statistical harmonization, we modeled between-site effects by using several different data handling strategies and investigated the association between relevant variables and classification accuracy, as described below.

Support vector machine classifier

We a priori chose to use support vector machine (SVM [50]), which is the most frequently used ML method in psychiatric brain imaging [15, 51]. The present analyses implemented a linear kernel, because this limits the risk of overfitting, contains only a single parameter, see below, and the coefficients of a linear classifier can be interpreted as relative measures of feature importance. However, we also performed sensitivity analyses to determine the impact of using a non-linear kernel (radial basis function) on results. All ML analyses were implemented in the Python programming language v. 3.6 using the Scikit-Learn package v. 0.19 [52].

The linear kernel SVM has only a single parameter, C, which controls the trade-off between having zero training errors and allowing misclassifications. We decided to a priori fix the hyperparameter at $C = 1$, for the following reasons. First, this setting is a common choice in the existing literature [53–56]. Second, SVM performance is relatively robust to changes in C values [57]. Third, the decision to perform hyperparameter optimization has data costs, as one must perform a further nesting of cross-validation, resulting in smaller effective training sets [58]. Also, hyperparameter optimization involves many steps, which have not been standardized and which may contribute to vibration of effects, including introduction of further hyperparameters (of the optimizers), selection of the best objective function over which to optimize, selection of constraints over the hyperparameter being optimized and of the hyperparameter optimization algorithm. Nevertheless, we also performed sensitivity analyses to determine the impact of hyperparameter optimization in a nested cross-validation procedure, see Supplementary material.

As the features used in the present study are engineered (i.e., the feature set does not consist of raw, voxelwise images), we opted against further feature selection. This decision was also supported by the large sample size and the fact that we had 20 times more participants than features. Importantly, in the above-described methods, the SVM models are independent across folds and no statistical harmonization, model selection or comparison was done prior to splitting the samples into testing and training.

Table 1 Descriptive statistics of the whole sample

	Controls	Cases	*p*-Value
N	2167	853	
Age mean (SD)	34.89 (12.41)	37.43 (11.64)	< 0.001
Sex, *N* (%) female	1201 (55.4)	516 (60.5)	0.013
Diagnosis, *N* (%)			
BD-I	-	582 (68.63)	
BD-II	-	234 (27.59)	
BD-NOS	-	13 (1.53)	
SZA	-	19 (2.24)	
Treatment at the time of scanning, *N* (%)			
Li		265 (33.5)	
AED	-	339 (43.1)	
FGA	-	32 (4.1)	
SGA	-	313 (39.9)	
AD	-	281 (35.5)	
Mood state, *N* (%)			
Euthymic	-	475 (75.5)	
Depressed	-	131 (20.8)	
Manic	-	11 (1.7)	
Hypomanic	-	9 (1.4)	
Mixed	-	3 (0.5)	
Age of onset mean (SD)	-	22.36 (9.08)	
Duration of illness mean (SD)	-	14.64 (10.45)	
History of psychosis, *N* (%)	-	372 (61.1)	

AD antidepressants, *AED* antiepileptics, *BD-I* bipolar I disorder, *BD-II* bipolar II disorder, *BD-NOS* bipolar disorder not otherwise specified, *FGA* first-generation antipsychotics, *Li* lithium, *SD* standard deviation, *SGA* second-generation antipsychotics, *SZA* schizoaffective disorder

Consequently, we have minimized the potential for information leak.

Classification performance was measured using standard metrics including accuracy, sensitivity, specificity, positive predictive value, negative predictive value and area under the receiver operating characteristic curve (ROC-AUC).

Data handling

The first application of the above-described classifier was to the classification of cases versus controls in individual sites, referred to as site-level analyses. For each site, we fit an SVM and measured its performance using a stratified K-fold cross-validation procedure. This method is stratified insofar as the proportion of cases and controls (in respective folds) is similar in both training and validation sets. The number of folds was selected independently for each site, such that the validation set on each fold would have approximately 3 (±1) cases.

To further study how overall classification performance relates to different methods of data handling, we implemented three approaches. The first was a meta-analysis of diagnostic accuracy from site-level analyses, referred to as meta-analysis. This models the typical method of analyzing data in a multi-site collaboration [11, 14]. The meta-analyses were done using the hierarchical summary receiver operating characteristic, implemented in HSROC package v. 2.1.8 [59], in the R programming language, see Supplementary material.

Second, we evaluated the same linear SVM parameterization used in all other analyses on a leave-one-site-out (LOSO) cross-validation procedure, referred to as LOSO analyses. In each fold of cross-validation, one site's data were completely excluded from the training partition. The SVM was then trained on the training partition and predictive performance was evaluated on the data from the held-out site.

Third, we fit an SVM classifier to the data pooled across all sites, using the same linear SVM parameterization as in the site-level analyses, and the same cross-validation procedure. This yielded a total of 284-folds and is further referred to as aggregate subject-level analysis.

We corrected for the effects of imbalanced data in all analyses and thereby trained the SVM classifiers on an effectively balanced dataset. To do this, we implemented the Synthetic Minority Oversampling Technique with Tomek link [60, 61] using the imblearn package v. 0.3.0.dev0 [62], in the Python language v. 3.6. The computer code for the above-described analyses will be provided upon reasonable request.

Feature importance

To determine feature importance, we plotted the SVM coefficients learned (over a total of K-folds per sample) based on the aggregated data and the SVM coefficients learned from the site with the highest ROC-AUC performance. To quantify whether similar features contributed to classification in different analyses, we computed Cohen's kappa for agreement in ranking of feature coefficients of individual regions between these two models, see Supplementary material for details of this calculation.

Investigation of clinical heterogeneity/potential confounding factors

We investigated whether any confounding factors contributed to the classification by examining the relationship between clinical/demographic variables and classification results using mixed-effects logistic regression – glmer function in the lme4 package of the R Statistical Programming Language [63]. Variables listed in Table 1 and

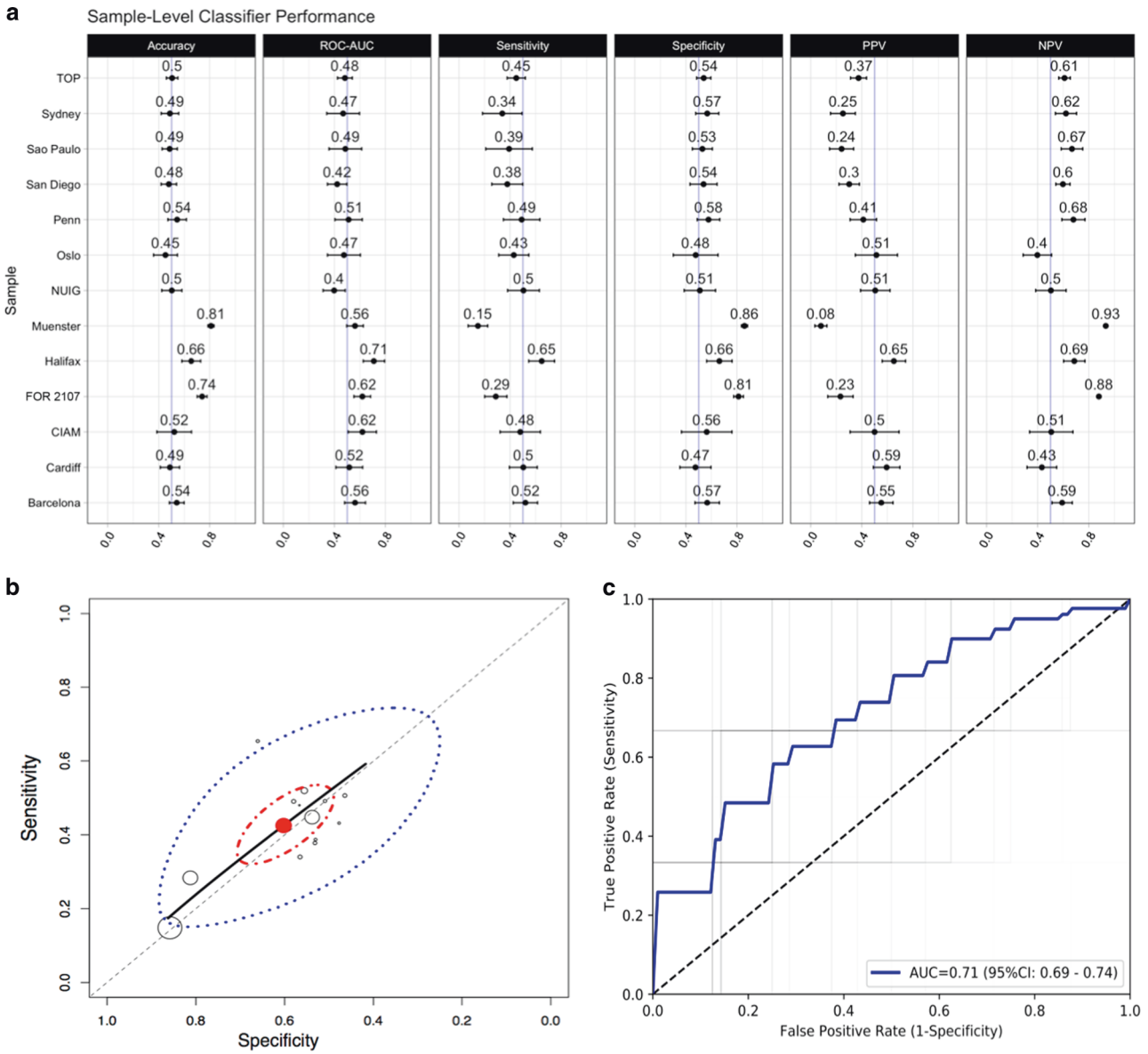


Fig. 1 **a** Performance of SVM classifiers independently trained on each sample – mean with 95% confidence interval. Each row denotes a site in the data set, whereas each column denotes a specific performance metric. **b** Meta-analytic (summary) receiver operating characteristic (SROC) curves. Site-level sensitivity (Sn) and specificity (Sp) are empty circles of radius proportional to sample size. The red point is the median estimate of Sn and Sp. The solid black line is the SROC curve. Dashed diagonal represents chance performance. The red ellipse is the 95% posterior credible region, and the blue dashed line is the 95% posterior predictive region. **c** Receiver operating characteristic (ROC) curves for the aggregate subject-level analysis. Faint gray lines are the ROC curves for individual validation folds, and blue lines represent the mean ROC curve

intercepts were taken as random effects varying between sites about a group mean, see Supplementary material. For numerical stability, age, age of onset and duration of illness were scaled to have mean 0 and unit variance.

Results

We included 3020 participants (853 BD cases and 2167 controls), see Table 1.

The classification accuracy in individual sites ranged from 45.23% (95% confidence interval (95% CI) = 35.91–54.57) to 81.07% (95% CI = 78.68–83.46), see Fig. 1a. The classification performance was closely associated with the method of data handling. Meta-analysis of individual site results yielded the lowest performance, which did not exceed chance level, see Fig. 1b, Table 2. The LOSO cross-validation provided above chance classification, but performed worse than the aggregate subject-level analyses. Aggregating the data across sites yielded the highest and

Table 2 Summary of classification results from meta-analysis of sample-level classifiers, leave-one-site-out and aggregate subject-level analyses

Statistic	Meta-analysis	Leave-one-site-out	Aggregate subject-level
Accuracy (%)	-	58.67 (56.70–60.63)	65.23 (63.47–67.00)
ROC-AUC	-	60.92 (58.18–63.67)	71.49 (69.39–73.59)
Sensitivity (%)	42.60 (13.40–71.57)	51.99 (48.20–55.78)	66.02 (62.71–69.33)
Specificity (%)	59.14 (30.59–87.94)	64.85 (61.91–67.79)	64.90 (62.86–66.93)
PPV (%)	-	47.25 (37.67–56.84)	44.45 (42.04–46.86)
NPV (%)	-	67.67 (60.36–74.98)	83.73 (82.21–85.26)

Note that meta-analytic results of the HSROC package include only sensitivity and specificity of the overall meta-analytic classification. Results for meta-analytic summary are the posterior predictive value of the performance metric, reported as mean (95% credible interval; the Bayesian analog of 95% confidence intervals). Results for the aggregate subject-level and leave-one-site-out analyses are reported as mean and 95% confidence interval

NPV negative predictive value, *PPV* positive predictive value, *ROC-AUC* area under receiver operating characteristic curve

statistically significant classification performance, see Fig. 1c, Table 2.

Feature importance

Ranking of features, which contributed to classification in the site with the highest ROC-AUC and the aggregate subject-level analyses, see Fig. 2, showed substantial agreement (Cohen's Kappa = 0.83, 95% CI = 0.829–0.831).

Effects of clinical heterogeneity

Among BD participants in the aggregate subject-level analysis, both age (odds ratio (OR) = 1.4, 95% CI = 1.05–1.88, $p = 0.02$) and antiepileptic use (OR = 1.73, 95% CI = 1.07–2.78, $p = 0.02$) were positively and additively associated with correct classification. There was no association between correct classification and diagnostic subgroup, treatment with first-, second-generation antipsychotics, lithium (Li), age of onset, history of psychosis, mood state or sex, see Supplementary Table S5. Age was necessarily co-linear with duration of illness ($r(782) = 0.66$, $p < 0.001$), but there was no univariate association between the duration of illness and correct classification (OR = 1.18, 95% CI = 0.98–1.43, $p = 0.09$).

Treatment with anticonvulsants was negatively associated with Li treatment (OR = 0.39, 95% CI = 0.19–0.80, $p = 0.01$), but not with any other clinical features, see Supplementary Table S6.

In the whole sample, both age (OR = 1.46, 95% CI = 1.17–1.81, $p < 0.001$) and status (BD versus controls; OR = 1.60, 95% CI = 1.28–2.01, $p < 0.001$), but not sex (OR = 1.21, 95% CI = 0.99−1.48, p=0.06) were independently associated with being classified as a BD participant.

Sensitivity analyses

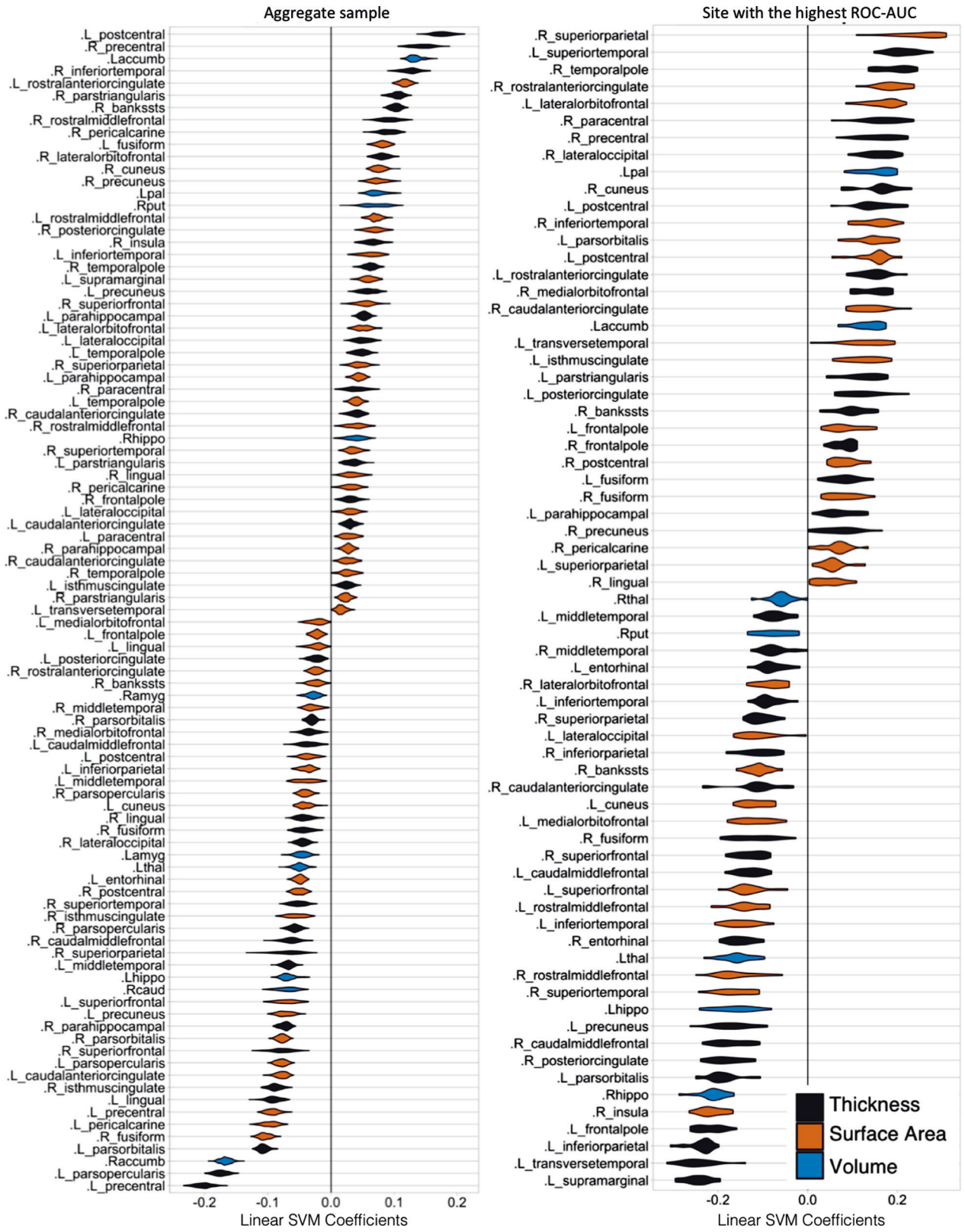
Using the radial basis function kernel yielded accuracy of 68%, 95% CI = 67–69%. Hyperparameter optimization resulted in training set accuracy of 65.9%, 95% CI = 65.7–66.0 and testing set accuracy of 57.5%, 95% CI = 49.1–65.9. Thus, it is unlikely that substantial classification performance was sacrificed by forgoing kernel nonlinearity or hyperparameter optimization.

In the LOSO analysis, when we left out the sites with the highest ROC-AUC curves, i.e., Halifax, Marburg (FOR 2107), Cape Town (CIAM), we acquired ROC-AUC of 65.42%, 66.18%, 63.07%, respectively, see Fig. 3, which was comparable to the overall ROC-AUC of 60.92% in the LOSO. Thus, the overall results did not appear to be overly influenced by the best performing sites.

Discussion

When applied to structural brain-imaging data, ML differentiated BD participants from controls with above chance accuracy even in a large and heterogeneous sample of 3020 participants from 13 sites worldwide. Aggregate analyses of individual subject data yielded better performance than LOSO or meta-analysis of site-level results. Despite the multi-site nature, ML identified a set of plausible brain-imaging features, which characterized individual BD participants and generalized across samples. Age and exposure to anticonvulsants were associated with greater odds of correct classification.

Previous studies employing raw structural MRI data have yielded accuracies between 59.50 and 73.00% [22, 23] for differentiating BD from control participants. A single study using results from automated segmentation reported accuracy below 60.00% [37]. Although direct


Aggregate sample
.L_postcentral
.R_precentral
.Laccumb
.R_inferiortemporal
.L_rostralanteriorcingulate
.R_parstriangularis
.R_bankssts
.R_rostralmiddlefrontal
.R_pericalcarine
.L_fusiform
.R_lateralorbitofrontal
.R_cuneus
.R_precuneus
.Lpal
.Rput
.L_rostralmiddlefrontal
.R_posteriorcingulate
.R_insula
.L_inferiortemporal
.R_temporalpole
.L_supramarginal
.L_precuneus
.R_superiorfrontal
.L_parahippocampal
.L_lateralorbitofrontal
.L_lateraloccipital
.L_temporalpole
.R_superiorparietal
.L_parahippocampal
.R_paracentral
.L_temporalpole
.R_caudalanteriorcingulate
.R_rostralmiddlefrontal
.Rhippo
.R_superiortemporal
.L_parstriangularis
.R_lingual
.R_pericalcarine
.R_frontalpole
.L_lateraloccipital
.L_caudalanteriorcingulate
.L_paracentral
.R_parahippocampal
.R_caudalanteriorcingulate
.R_temporalpole
.L_isthmuscingulate
.R_parstriangularis
.L_transversetemporal
.L_medialorbitofrontal
.L_frontalpole
.L_lingual
.L_posteriorcingulate
.R_rostralanteriorcingulate
.R_bankssts
.Ramyg
.R_middletemporal
.R_parsorbitalis
.R_medialorbitofrontal
.L_caudalmiddlefrontal
.L_postcentral
.L_inferiorparietal
.L_middletemporal
.R_parsopercularis
.L_cuneus
.R_lingual
.R_fusiform
.R_lateraloccipital
.Lamyg
.Lthal
.L_entorhinal
.R_postcentral
.R_superiortemporal
.R_isthmuscingulate
.R_parsopercularis
.R_caudalmiddlefrontal
.R_superiorparietal
.L_middletemporal
.Lhippo
.Rcaud
.L_superiorfrontal
.L_precuneus
.R_parahippocampal
.R_parsorbitalis
.R_superiorfrontal
.L_parsopercularis
.L_caudalanteriorcingulate
.R_isthmuscingulate
.L_lingual
.L_precentral
.L_pericalcarine
.R_fusiform
.L_parsorbitalis
.Raccumb
.L_parsopercularis
.L_precentral
-0.2
-0.1
0.0
0.1
0.2
Linear SVM Coefficients
Site with the highest ROC-AUC
.R_superiorparietal
.L_superiortemporal
.R_temporalpole
.R_rostralanteriorcingulate
.L_lateralorbitofrontal
.R_paracentral
.R_precentral
.R_lateraloccipital
.Lpal
.R_cuneus
.L_postcentral
.R_inferiortemporal
.L_parsorbitalis
.L_postcentral
.L_rostralanteriorcingulate
.R_medialorbitofrontal
.R_caudalanteriorcingulate
.Laccumb
.L_transversetemporal
.L_isthmuscingulate
.L_parstriangularis
.L_posteriorcingulate
.R_bankssts
.L_frontalpole
.R_frontalpole
.R_postcentral
.L_fusiform
.R_fusiform
.L_parahippocampal
.R_precuneus
.R_pericalcarine
.L_superiorparietal
.R_lingual
.Rthal
.L_middletemporal
.Rput
.R_middletemporal
.L_entorhinal
.R_lateralorbitofrontal
.L_inferiortemporal
.R_superiorparietal
.L_lateraloccipital
.R_inferiorparietal
.R_bankssts
.R_caudalanteriorcingulate
.L_cuneus
.L_medialorbitofrontal
.R_fusiform
.R_superiorfrontal
.L_caudalmiddlefrontal
.L_superiorfrontal
.L_rostralmiddlefrontal
.L_inferiortemporal
.R_entorhinal
.Lthal
.R_rostralmiddlefrontal
.R_superiortemporal
.Lhippo
.L_precuneus
.R_caudalmiddlefrontal
.R_posteriorcingulate
.L_parsorbitalis
.Rhippo
.R_insula
.L_frontalpole
.L_inferiorparietal
.L_transversetemporal
.L_supramarginal
Thickness
Surface Area
Volume
-0.2
0.0
0.2
Linear SVM Coefficients

◀ **Fig. 2** Violin plot of feature importance across cross-validation (CV) folds for aggregate subject-level analysis (left), and the site, which yielded the highest ROC-AUC (right). At each CV iteration, we extracted linear support vector machine (SVM) coefficients. The set of all coefficients from our SVM models are centered about 0. Deviation of coefficients from zero is an indication of the relative importance of individual features in the data. Features with positive and negative coefficients have positive and negative associations, respectively, with probability of classification as a case. The *y* axis lists variables for which SVM coefficients were strictly non-zero throughout all cross-validation iterations

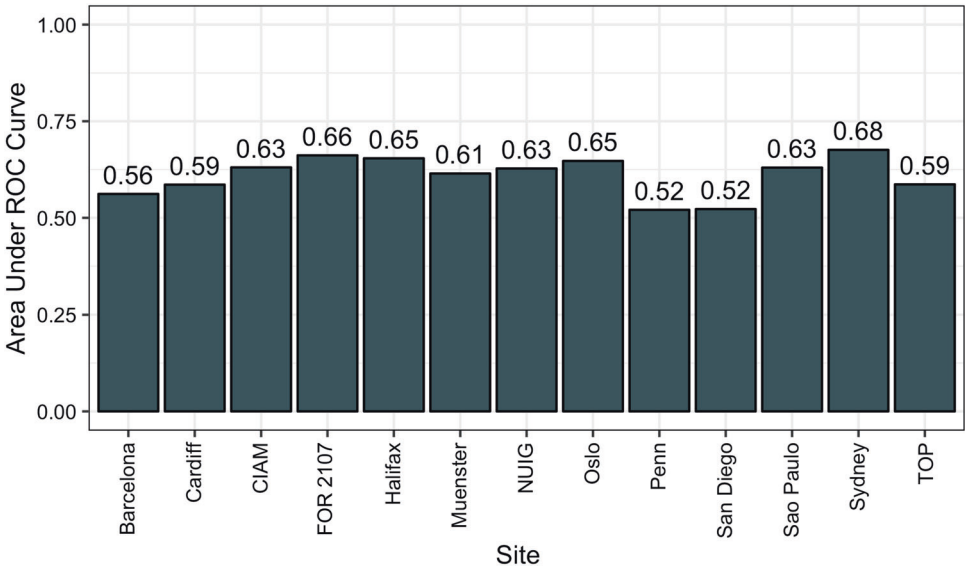


Fig. 3 Bar plot of the area under the receiver operating characteristic curve (ROC-AUC) for the leave-one-site-out (LOSO) analyses. The sites listed along the *x* axis are those that were held-out at each fold

comparison is complicated by methodological and sample size differences, the modest accuracies in previous studies are in keeping with our results. Thus, the presented findings appear realistic and there is little evidence for overfitting.

The classification performance in the aggregated dataset was significantly above chance level and the ROC-AUC of 71.49% (69.39–73.59) reached acceptable discrimination [64, 65]. However, the accuracy of 65.23% (95% CI = 63.47–67.00) fell short of the 80% threshold, which is deemed clinically relevant [66]. We need to consider several issues when interpreting these findings. BDs are difficult to diagnose even by standard methods. The Cohen's kappa for reliability of the BD-I diagnosis is 0.56 and as low as 0.40 for BD-II [67]. In addition, the illness shows marked clinical and neurobiological heterogeneity [10, 12]. Perhaps most importantly, we worked with regional brain measures, not raw/voxelwise data. This approach necessarily involves some information loss in the feature engineering process. Analyses of experimenter-defined features are increasingly outperformed by models capable of learning abstractions from raw data alone [68]. Applying deep learning [69] to raw data would likely offer the greatest increase in classification accuracy.

This study provides important clues about the impact of data handling on the classification performance. As expected, the meta-analysis of individual site results, typically the first method of data analyses in multi-site collaborations, yielded the lowest accuracy, which did not exceed chance level. The LOSO analyses performed better than the meta-analytic approach, but worse than when individual subject data were aggregated and analyzed jointly. These differences in performance are likely related to the way each method handles the conditional relationships between the sites. Meta-analyses essentially model these relationships after the fact. The LOSO analyses are hindered by the fact that data are partitioned along some factor that is not random. In contrast, pooling of data allows for random partitioning and incorporates the relationships between the sites in their raw form. In addition, the classification performance is closely linked to the size of the training sample [49, 70], which increased from individual site through LOSO to aggregate analyses.

Thus, the empirical pattern of findings is convergent with theoretical prediction of how each of these methods should perform. It is also congruent with previous studies in autism [49], schizophrenia [70] and Alzheimer dementia [26], which also showed increasing performance with increasing size of the training set. It is a question whether this would also be the case in more heterogeneous conditions, such as major depression or anxiety disorders. Regardless, aggregate analyses provided the best classification performance in BD. Future multi-site brain-imaging studies should attempt to move towards sharing of individual subject data, not only site-level results.

The linear SVM kernel allowed us to visualize the contribution of individual regions to the overall classification. It

is of note that the results of a backward model should not be used for localization [71]. We used this approach to broadly verify the neurobiological plausibility [26], not to infer pathophysiology. Our findings showed good validity, as many of the same regions, which have previously shown differences between groups of BD patients and controls, contributed to the classification on individual subject level, including hippocampus, amygdala [9–11], as well as cortical regions, such as inferior frontal gyrus [12, 14] and precentral gyrus [13].

In addition, we wanted to determine whether similar features were used for classification across different analyses. Indeed, there was a substantial agreement between the regions, which contributed to the classification in the site, which yielded the highest ROC-AUC and in the aggregate dataset, with Cohen's Kappa of 0.83 (95% CI = 0.829–0.831). Furthermore, when we trained the classifier on data from all but the best performing sites, the classification performance did not drop below the overall accuracy in the LOSO analyses. Thus, individual sites did not markedly influence the overall findings. Taken together, these results suggest that the classification was based on a biologically plausible and generalizable neurostructural signature, which is shared among subjects in a large, multi-site sample. This is highly interesting, as existence of a generalizable biomarker is one of the key defining features of a diagnostic category [72].

We also investigated the effects of clinical/demographic variables on classification accuracy. Older age and anticonvulsant treatment were associated with greater odds of correct classification. The effect of age may be related to the fact that illness-related alterations may get worse with age/duration of illness [73]. Interestingly, similar association was noted in a meta-analysis of brain-imaging ML studies in schizophrenia [74]. These findings also broadly agree with another study, in which late-stage BD was easier to classify than early stage illness [36]. However, we did not find an association between accuracy of classification and duration of illness or age of onset.

The association with anticonvulsant treatment may reflect effects of illness or medications. Treatment with anticonvulsants was not associated with severity of illness, diagnostic category, mood state, age of onset or personal history of psychotic symptoms and thus did not appear to index a specific subgroup within BD. Interestingly, participants who were treated with anticonvulsants were less likely to also receive Li treatment. Perhaps, the neuroprotective effects of Li, which may normalize brain alterations in BD [10, 75] could presumably make the classification based on brain structure more difficult. However, Li treatment itself was not associated with classification accuracy. Previous studies have suggested that valproate, may negatively affect brain structure [76], which could contribute to correct differentiation of anticonvulsant treated from control participants. This was, however, not documented for lamotrigine, which is also frequently used in treatment of BD. Overall, the reasons why treatment with anticonvulsants and age were associated with easier classification are unclear and will be subject to future analyses.

A related question is whether the clinical/demographic heterogeneity confounded our findings and whether the age and/or treatment with anticonvulsants contributed more to the classification than the presence or absence of BD. Due to selection bias, heterogeneity is more likely to affect results in smaller studies [25]. The strength of a large, multicenter analysis is that it will primarily target the common alterations, which are generalizable to most participants and not individual sources of heterogeneity, which are present only in some [25]. In addition, both age and status were independently and additively associated with being classified as a BD participant in the whole sample. Also, within the site with the highest classification performance, BD participants and controls were balanced by age. In addition, 43.1% of patients in the whole sample were treated with anticonvulsants and yet, we reached a 66.02% sensitivity for correctly identifying BD participants. Last but not least, the sites with the highest proportion of anticonvulsant-treated participants (61.4%) and the highest discrepancy in age showed relatively low sensitivities of 49% and 29%, respectively. Thus, although certain clinical and demographic variables were associated with correct classification, it is unlikely that overall we were classifying participants based on the presence or absence of specific clinical/demographic variables, rather than the presence or absence of BD.

Our study has the following limitations. Due to differences in availability, we did not include other brain-imaging modalities or other types of data, that is, genetic, neurocognitive or biochemical. Access to raw data would allow us to use deep learning methods [68] or create a meta-model by combining classifiers trained on the local datasets [77]. However, currently there are significant practical and legal limitations to raw data sharing. The clinical heterogeneity and multi-site nature, which complicate traditional between-group comparisons, allowed us to test the ML algorithms on a wide range of participants in a fair setting that better resembles a clinical situation. To achieve a clearer exposition and reduce methodological heterogeneity, we decided to use SVM. Previous studies have generally found minimal differences between "shallow" ML method [37]. As we worked with regional brain measures, not voxelwise data, we would not be able to fully exploit the power of deeper methods [78]. The depth and breadth of phenotyping are general issues in retrospective, multi-site data sharing collaborations. Specific sources of heterogeneity, that is, neuroprogression and comorbid conditions, may be particularly

difficult to quantify. Addressing them would require a different research design. However, the large, multi-site sample, together with the exploratory analyses and examination of individual site results made it less likely that individual clinical characteristics systematically confounded the findings. Finally, attempting to differentiate BD from control participants is the first step before moving to more clinically relevant problems, such as differential diagnosis.

The key advantages of this study include the large, generalizable sample, access to individual subject data from 13 sites and the conservative and scalable nature of the analyses. This is currently the largest application of ML to brain-imaging data in BD, with up to two orders of magnitude, greater sample size than in previous studies. The unique nature of the dataset provides qualitative, not only quantitative advantages. Previous studies showed low stability of ML results with fewer than 130 participants [70], a threshold we exceeded 7–16 times. The multi-site dataset maximized the training set size, provided ecologically valid representation of the illness, allowed us to focus on common, BD-related alterations and for the first time apply the LOSO cross-validation in BD brain imaging. We completely separated the testing and training sets at each level of analysis, thus minimizing the risk of information leak, and specifically focused on maximizing generalizability/reducing the risk of overfitting. The study is an example of close international collaboration, which is one of the best ways, how to create optimal datasets for ML analyses.

Conclusions

This study provides a realistic and fair estimate of classification performance, which can be achieved in a large, ecologically valid, multi-site sample of BD participants based on regional neurostructural measures. Although short of the 80% clinically relevant threshold the 65.23% accuracy, 71.49% ROC-AUC are promising, as we used an engineered feature set in a difficult to diagnose condition, which shows a marked clinical and neurobiological heterogeneity. In addition, similar, biologically plausible features contributed to classification in different analyses. Together these findings provide a proof of concept for a generalizable brain-imaging signature of BD, which can be detected on individual subject level, even in a large, multi-site sample. Although specific clinical/demographic characteristics, such age and anticonvulsant treatment, may affect classification, the clinical heterogeneity was not in the way of differentiating BD from control participants. Finally, we demonstrated that meta-analyses of individual site/study ML performances provide a poor proxy for results, which could be obtained by pooling of individual subject data. These findings are an important step towards translating brain imaging from bench to the bedside. They suggest that a multi-site ML classifier may correctly identify previously unseen data and aid in diagnosing individual BD participants. Application of deep learning to raw data might considerably increase the accuracy of classification.

Acknowledgements The researchers and studies included in this paper were supported by Australian National Medical and Health Research Council (Program Grant 1037196) and the Lansdowne Foundation, Generalitat de Catalunya PERIS postdoc contract (Plaestrategic de Recerca i Innovacio en Salut), the Spanish Ministry of Economy and Competitiveness (PI15/00283) integrated into the Plan Nacional de I + D + I y cofinanciado por el ISCIII-Subdirección General de Evaluación y el Fondo Europeo de Desarrollo Regional (FEDER); CIBERSAM; and the Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya to the Bipolar Disorders Group (2014 SGR 398), Dalhousie Department of Psychiatry and Clinician Investigator Program (AN), grants of the Deanery of the Medical Faculty of the University of Münster, National Institute of Mental Health grant MH083968, R01MH101111, R01MH107703, K23MH85096; Desert-Pacific Mental Illness Research, Education, and Clinical Center, 2010 NARSAD Young Investigator Award to Dr Xavier Caseras, PRISMA U.T., Colciencias, Research Council of Norway (249795), the South-Eastern Norway Regional Health Authority (2014097), Research grant Health Region South-East, Norway and Throne-Holst research grant, South African Medical Research Council, Australian National Medical and Health Research Council (NHMRC) Program Grant 1037196, 1063960 and 1066177, Janette Mary O'Neil Research Fellowship, the European Commission's 7th Framework Programme #602450 (IMAGEMEND), #602450, the FIDMAG Hermanas Hospitalarias Research Foundation sample is supported by the Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya (2017-SGR-1271) and several grants funded by Instituto de Salud Carlos III (Co-funded by European Regional Development Fund/European Social Fund) "Investing in your future"): Miguel Servet Research Contract (CPII16/00018 to EP-C and CPII13/00018 to RS), Sara Borrell Contract grant (CD16/00264 to MF-V) and Research Projects (PI14/01148 to EP-C, PI14/01151 to RS and PI15/00277 to EJC-R), Health Research Board (grant number HRA_POR/2011/100), Agence Nationale pour la Recherche (ANR-11-IDEX-0004 Labex BioPsy, ANR-10-COHO-10-01 psyCOH), Fondation pour la Recherche Médicale (Bioinformatique pour la biologie 2014) and the Fondation de l'Avenir (Recherche Médicale Appliquée 2014), The Research Council of Norway (#223273, #229129, #249711), KG Jebsen Stiftelsen (SKGJ-MED-008), the South-Eastern Norway Regional Health Authority, Oslo University Hospital, the Ebbe Frøland foundation, and a research grant from Mrs. Throne-Holst, FAPESP-Brazil 2013/03905-4), CNPq-Brazil (#478466/2009 and 480370/2009), and the Brain & Behavior Research Foundation (2010 NARSAD Independent Investigator Award to GFB), Innovative Medizinische Forschung (RE111604 to RR and RE111722 to RR); SFB-TRR58, Projects C09 and Z02 to UD and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/012/17 to UD), The German Research Foundation (DFG) as part of the Research Unit "Neurobiology of Affective Disorders" (DFG FOR 2107; KI 588/14-1, KO 4291/3-1, KI 588/14-1, KR 3822/5-1, DA 1151/5-1, DA1151/5-2, JA 1890/7-1, AK 3822/5-1), University Research Committee, University of Cape Town, South Africa; National Research Foundation, South Africa, Canadian Institutes of Health Research (103703, 106469 and 64410), Nova Scotia Health Research Foundation, Dalhousie Clinical Research Scholarship to T Hajek, NARSAD 2007 Young Investigator and 2015 Independent Investigator Awards to T Hajek, the NIH grant (U54 EB020403) to the ENIGMA Center for Worldwide Medicine,

Imaging & Genomics, funded as part of the NIH Big Data to Knowledge (BD2K) initiative.

Compliance with ethical standards

Conflict of interest OAA received speaker's honorarium from Lundbeck. JCS has participated in research funded by BMS, Forest, Merck, Elan, J&J consulted for Astellas and has been a speaker for Pfizer, Abbott and Sanofi. TE has received honoraria for lecturing from GlaxoSmithKlein, Pfizer, and Lundbeck. EV has received grants and served as consultant, advisor or speaker for the following entities: AB-Biotics, Allergan, Angelini, Dainippon Sumitomo Pharma, Farmindustria, Ferrer, Gedeon Richter, Janssen, Johnson and Johnson, Lundbeck, Otsuka, Pfizer, Roche, Sanofi-Aventis, Servier, the Brain and Behavior Foundation, the Seventh European Framework Programme (ENBREC), the Stanley Medical Research Institute, Sunovion, and Takeda. The remaining authors declare that they have no conflict of interest.

References

1. Gustavsson A, Svensson M, Jacobi F, Allgulander C, Alonso J, Beghi E, et al. Cost of disorders of the brain in Europe 2010. Eur Neuropsychopharmacol. 2011;21:718–79.
2. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet. 2013;382:1575–86.
3. Bschor T, Angst J, Azorin JM, Bowden CL, Perugi G, Vieta E, et al. Are bipolar disorders underdiagnosed in patients with depressive episodes? Results of the multicenter BRIDGE screening study in Germany. J Affect Disord. 2012;142:45–52.
4. Ghaemi SN, Sachs GS, Chiou AM, Pandurangi AK, Goodwin K. Is bipolar disorder still underdiagnosed? Are antidepressants overutilized? J Affect Disord. 1999;52:135–44.
5. Duffy A, Alda M, Hajek T, Grof P. Early course of bipolar disorder in high-risk offspring: prospective study. Br J Psychiatry. 2009;195:457–8.
6. Conus P, Macneil C, McGorry PD. Public health significance of bipolar disorder: implications for early intervention and prevention. Bipolar Disord. 2014;16:548–56.
7. Schmitt A, Rujescu D, Gawlik M, Hasan A, Hashimoto K, Iceta S, et al. Consensus paper of the WFSBP Task Force on Biological Markers: criteria for biomarkers and endophenotypes of schizophrenia part II: cognition, neuroimaging and genetics. World J Biol Psychiatry. 2016;17:406–28.
8. Woodcock J, Woosley R. The FDA critical path initiative and its influence on new drug development. Annu Rev Med. 2008;59:1–12.
9. Hajek T, Kopecek M, Kozeny J, Gunde E, Alda M, Hoschl C. Amygdala volumes in mood disorders - meta-analysis of magnetic resonance volumetry studies. J Affect Disord. 2009;115:395–410.
10. Hajek T, Kopecek M, Hoschl C, Alda M. Smaller hippocampal volumes in patients with bipolar disorder are masked by exposure to lithium: a meta-analysis. J Psychiatry Neurosci. 2012;37:110143.
11. Hibar DP, Westlye LT, van Erp TG, Rasmussen J, Leonardo CD, Faskowitz J, et al. Subcortical volumetric abnormalities in bipolar disorder. Mol Psychiatry. 2016;21:1710–6.
12. Hajek T, Cullis J, Novak T, Kopecek M, Blagdon R, Propper L, et al. Brain structural signature of familial predisposition for bipolar disorder: replicable evidence for involvement of the right inferior frontal gyrus. Biol Psychiatry. 2013;73:144–52.
13. Ganzola R, Duchesne S. Voxel-based morphometry meta-analysis of gray and white matter finds significant areas of differences in bipolar patients from healthy controls. Bipolar Disord. 2017;19:74–83.
14. Hibar DP, Westlye LT, Doan NT, Jahanshad N, Cheung JW, Ching CRK, et al. Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. Mol Psychiatry. 2018;23:932–42.
15. Orru G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci Biobehav Rev. 2012;36:1140–52.
16. Fu CH, Costafreda SG. Neuroimaging-based biomarkers in psychiatry: clinical opportunities of a paradigm shift. Can J Psychiatry. 2013;58:499–508.
17. Davatzikos C. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. Neuroimage. 2004;23:17–20.
18. Castellanos FX, Di Martino A, Craddock RC, Mehta AD, Milham MP. Clinical applications of the functional connectome. Neuroimage. 2013;80:527–40.
19. Milham MP, Craddock RC, Klein A. Clinically useful brain imaging for neuropsychiatry: how can we get there? Depress Anxiety. 2017;34:578–87.
20. Atluri G, Padmanabhan K, Fang G, Steinbach M, Petrella JR, Lim K, et al. Complex biomarker discovery in neuroimaging data: finding a needle in a haystack. Neuroimage Clin. 2013;3:123–31.
21. Davatzikos C, Shen D, Gur RC, Wu X, Liu D, Fan Y, et al. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. Arch Gen Psychiatry. 2005;62:1218–27.
22. Schnack HG, Nieuwenhuis M, van Haren NE, Abramovic L, Scheewe TW, Brouwer RM, et al. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. Neuroimage. 2014;84:299–306.
23. Rocha-Rego V, Jogia J, Marquand AF, Mourao-Miranda J, Simmons A, Frangou S. Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach. Psychol Med. 2014;44:519–32.
24. Bansal R, Staib LH, Laine AF, Hao X, Xu D, Liu J, et al. Anatomical brain images alone can accurately diagnose chronic neuropsychiatric illnesses. PLoS ONE. 2012;7:e50698.
25. Schnack HG, Kahn RS. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. Front Psychiatry. 2016;7:50.
26. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci. 2017;20:365–77.
27. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage. 2017;S1053-8119:30531-1.

28. Mwangi B, Spiker D, Zunta-Soares GB, Soares JC. Prediction of pediatric bipolar disorder using neuroanatomical signatures of the amygdala. Bipolar Disord. 2014;16:713–21.
29. Jie NF, Zhu MH, Ma XY, Osuch EA, Wammes M, Theberge J, et al. Discriminating bipolar disorder from major depression based on SVM-FoBa: efficient feature selection with multimodal brain imaging data. IEEE Trans Auton Ment Dev. 2015;7:320–31.
30. Serpa MH, Ou Y, Schaufelberger MS, Doshi J, Ferreira LK, Machado-Vieira R, et al. Neuroanatomical classification in a population-based sample of psychotic major depression and bipolar I disorder with 1 year of diagnostic stability. Biomed Res Int. 2014;2014:706157.
31. Fung G, Deng Y, Zhao Q, Li Z, Qu M, Li K, et al. Distinguishing bipolar and major depressive disorders by brain structural morphometry: a pilot study. BMC Psychiatry. 2015;15:298.
32. Rubin-Falcone H, Zanderigo F, Thapa-Chhetry B, Lan M, Miller JM, Sublette ME, et al. Pattern recognition of magnetic resonance imaging-based gray matter volume measurements classifies bipolar disorder and major depressive disorder. J Affect Disord. 2017;227:498–505.
33. Sacchet MD, Livermore EE, Iglesias JE, Glover GH, Gotlib IH. Subcortical volumes differentiate major depressive disorder, bipolar disorder, and remitted major depressive disorder. J Psychiatr Res. 2015;68:91–8.
34. Koutsouleris N, Meisenzahl EM, Borgwardt S, Riecher-Rossler A, Frodl T, Kambeitz J et al. Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. Brain 2015;138:2059–73.
35. Doan NT, Kaufmann T, Bettella F, Jorgensen KN, Brandt CL, Moberget T, et al. Distinct multivariate brain morphological patterns and their added predictive value with cognitive and polygenic risk scores in mental disorders. Neuroimage Clin. 2017;15:719–31.
36. Mwangi B, Wu MJ, Cao B, Passos IC, Lavagnino L, Keser Z, et al. Individualized prediction and clinical staging of bipolar disorders using neuroanatomical biomarkers. Biol Psychiatry Cogn Neurosci Neuroimaging. 2016;1:186–94.
37. Salvador R, Radua J, Canales-Rodriguez EJ, Solanes A, Sarro S, Goikolea JM, et al. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. PLoS ONE. 2017;12:e0175683.
38. Redlich R, Almeida JJ, Grotegerd D, Opel N, Kugel H, Heindel W, et al. Brain morphometric biomarkers distinguishing unipolar and bipolar depression. A voxel-based morphometry-pattern classification approach. JAMA Psychiatry. 2014;71:1222–30.
39. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med. 2016;46:2455–65.
40. Kempton MJ, Haldane M, Jogia J, Grasby PM, Collier D, Frangou S. Dissociable brain structural changes associated with predisposition, resilience, and disease expression in bipolar disorder. J Neurosci. 2009;29:10863–8.
41. Roberts G, Lenroot R, Frankland A, Yeung PK, Gale N, Wright A, et al. Abnormalities in left inferior frontal gyral thickness and parahippocampal gyral volume in young people at high genetic risk for bipolar disorder. Psychol Med. 2016;46:2083–96.
42. Hajek T, Cullis J, Novak T, Kopecek M, Hoschl C, Blagdon R, et al. Hippocampal volumes in bipolar disorders: opposing effects of illness burden and lithium treatment. Bipolar Disord. 2012;14:261–70.
43. Kelly S, Jahanshad N, Zalesky A, Kochunov P, Agartz I, Alloza C, et al. Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA Schizophrenia DTI Working Group. Mol Psychiatry 2017;23:1261–69.
44. Schmaal L, Hibar DP, Samann PG, Hall GB, Baune BT, Jahanshad N, et al. Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. Mol Psychiatry. 2017;22:900–9.
45. Panizzon MS, Fennema-Notestine C, Eyler LT, Jernigan TL, Prom-Wormley E, Neale M, et al. Distinct genetic influences on cortical surface area and cortical thickness. Cereb Cortex. 2009;19:2728–35.
46. Winkler AM, Kochunov P, Blangero J, Almasy L, Zilles K, Fox PT, et al. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. Neuroimage. 2010;53:1135–46.
47. Lin A, Ching CRK, Vajdi A, Sun D, Jonas RK, Jalbrzikowski M, et al. Mapping 22q11.2 gene dosage effects on brain morphometry. J Neurosci. 2017;37:6183–99.
48. Rozycki M, Satterthwaite TD, Koutsouleris N, Erus G, Doshi J, Wolf DH, et al. Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. Schizophr Bull. 2017.
49. Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, et al. Deriving reproducible biomarkers from multisite resting-state data: an autism-based example. Neuroimage. 2017;147:736–45.
50. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
51. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage. 2017;145(Pt B):137–65.
52. Pedregosa F, Varoquaux GI, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2012;12:2825–30.
53. Mourao-Miranda J, Reinders AA, Rocha-Rego V, Lappin J, Rondina J, Morgan C, et al. Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. Psychol Med. 2012;42:1037–47.
54. Ecker C, Rocha-Rego V, Johnston P, Mourao-Miranda J, Marquand A, Daly EM, et al. Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. Neuroimage. 2010;49:44–56.
55. Hajek T, Cooke C, Kopecek M, Novak T, Hoschl C, Alda M. Using structural MRI to identify individuals at genetic risk for bipolar disorders: a 2-cohort, machine learning study. J Psychiatry Neurosci. 2015;40:316–24.
56. Pettersson-Yeo W, Benetti S, Marquand AF, Dell'acqua F, Williams SC, Allen P, et al. Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. Psychol Med. 2013;43:2547–62.
57. LaConte S, Strother S, Cherkassky V, Anderson J, Hu X. Support vector machines for temporal classification of block design fMRI data. Neuroimage. 2005;26:317–29.
58. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci Biobehav Rev. 2015;57:328–49.
59. Rutter CM, Gatsonis C. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20:2865–84.
60. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
61. He H, Garcia E. Learning from imbalanced data sets. IEEE Trans Knowl data Eng. 2010;21:1263–4.

62. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18:1–5.
63. Bates D, MÃ chler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:51.
64. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken, New Jersey, USA: Wiley; 2013
65. Iniesta R, Hodgson K, Stahl D, Malki K, Maier W, Rietschel M, et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Sci Rep. 2018;8:5530.
66. Savitz JB, Rauch SL, Drevets WC. Clinical application of brain imaging for the diagnosis of mood disorders: the current state of play. Mol Psychiatry. 2013;18:528–39.
67. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry. 2013;170:59–70.
68. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35:1798–828.
69. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
70. Nieuwenhuis M, van Haren NE, Hulshoff Pol HE, Cahn W, Kahn RS, Schnack HG. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. Neuroimage. 2012;61:606–12.
71. Haufe S, Meinecke F, Gorgen K, Dahne S, Haynes JD, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage. 2014;87:96–110.
72. Robins E, Guze SB. Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. Am J Psychiatry. 1970;126:983–7.
73. Berk M, Kapczinski F, Andreazza AC, Dean OM, Giorlando F, Maes M, et al. Pathways underlying neuroprogression in bipolar disorder: focus on inflammation, oxidative stress and neurotrophic factors. Neurosci Biobehav Rev. 2011;35:804–17.
74. Kambeitz J, Kambeitz-Ilankovic L, Leucht S, Wood S, Davatzikos C, Malchow B, et al. Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. Neuropsychopharmacology. 2015;40:1742–51.
75. Hajek T, Weiner MW. Neuroprotective effects of lithium in human brain? Food for thought. Curr Alzheimer Res. 2016;13:862–72.
76. Tariot PN, Schneider LS, Cummings J, Thomas RG, Raman R, Jakimovich LJ, et al. Chronic divalproex sodium to attenuate agitation and clinical progression of Alzheimer disease. Arch Gen Psychiatry. 2011;68:853–61.
77. Dluhos P, Schwarz D, Cahn W, Van Haren N, Kahn R, Spaniel F, et al. Multi-center machine learning in imaging psychiatry: a meta-model approach. Neuroimage. 2017;155:10–24.
78. Goodfellow I, Bengio Y, Courville A. Chapter 5: Machine learning basics. In *Deep Learning*. Cambridge, MA, USA: MIT Press; 2016.

Affiliations

Abraham Nunes[1,2] · Hugo G. Schnack[3] · Christopher R. K. Ching[4,5] · Ingrid Agartz[6,7,8,9] · Theophilus N. Akudjedu[10] · Martin Alda[1] · Dag Alnæs[6,7] · Silvia Alonso-Lana[11,12] · Jochen Bauer[13] · Bernhard T. Baune[14] · Erlend Bøen[8] · Caterina del Mar Bonnin[15] · Geraldo F. Busatto[16,17] · Erick J. Canales-Rodríguez[11,12] · Dara M. Cannon[10] · Xavier Caseras[18] · Tiffany M. Chaim-Avancini[16,17] · Udo Dannlowski[19] · Ana M. Díaz-Zuluaga[20] · Bruno Dietsche[21] · Nhat Trung Doan[6,7] · Edouard Duchesnay[22] · Torbjørn Elvsåshagen[6,23] · Daniel Emden[19] · Lisa T. Eyler[24,25] · Mar Fatjó-Vilas[11,12,26] · Pauline Favre[22] · Sonya F. Foley[27] · Janice M. Fullerton[28,29] · David C. Glahn[30,31] · Jose M. Goikolea[15] · Dominik Grotegerd[19] · Tim Hahn[19] · Chantal Henry[32] · Derrek P. Hibar[5] · Josselin Houenou[22,33] · Fleur M. Howells[34,35] · Neda Jahanshad[5] · Tobias Kaufmann[6,7] · Joanne Kenney[10] · Tilo T. J. Kircher[21] · Axel Krug[21] · Trine V. Lagerberg[6] · Rhoshel K. Lenroot[36,37] · Carlos López-Jaramillo[20,38] · Rodrigo Machado-Vieira[16,39] · Ulrik F. Malt[40,41] · Colm McDonald[10] · Philip B. Mitchell[36,42] · Benson Mwangi[39] · Leila Nabulsi[10] · Nils Opel[19] · Bronwyn J. Overs[28] · Julian A. Pineda-Zapata[43] · Edith Pomarol-Clotet[11,12] · Ronny Redlich[19] · Gloria Roberts[36,42] · Pedro G. Rosa[16,17] · Raymond Salvador[11,12] · Theodore D. Satterthwaite[44] · Jair C. Soares[39] · Dan J. Stein[45] · Henk S. Temmingh[45,46] · Thomas Trappenberg[2] · Anne Uhlmann[45,47] · Neeltje E. M. van Haren[3,48] · Eduard Vieta[15] · Lars T. Westlye[6,7,49] · Daniel H. Wolf[44] · Dilara Yüksel[21] · Marcus V. Zanetti[16,17,50] · Ole A. Andreassen[6,7] · Paul M. Thompson[5] · Tomas Hajek[1] for the ENIGMA Bipolar Disorders Working Group

[1] Department of Psychiatry, Dalhousie University, Halifax, Nova Scotia, Canada

[2] Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

[3] Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[4] Interdepartmental Neuroscience Program, University of California, Los Angeles, CA, USA

[5] Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Marina del Rey, CA, USA

[6] NORMENT KG Jebsen Centre, University of Oslo, Oslo, Norway

[7] Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

[8] Department of Psychiatric Research, Diakonhjemmet Hospital, Oslo, Norway

[9] Department of Clinical Neuroscience, Centre for Psychiatric Research, Karolinska Institutet, Stockholm, Sweden

[10] Centre for Neuroimaging and Cognitive Genomics (NICOG), Clinical Neuroimaging Laboratory, NCBES Galway Neuroscience Centre, College of Medicine Nursing and Health Sciences, National University of Ireland Galway, Galway, Ireland

[11] FIDMAG Germanes Hospitalàries Research Foundation, Barcelona, Spain

[12] Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid, Spain

[13] Institute of Clinical Radiology, Medical Faculty – University of Muenster – and University Hospital Muenster, Muenster, Germany

[14] Department of Psychiatry, Melbourne Medical School, The University of Melbourne, Parkville, VIC, Australia

[15] Hospital Clinic, University of Barcelona, IDIBAPS, CIBERSAM, Barcelona, Catalonia, Spain

[16] Laboratory of Psychiatric Neuroimaging (LIM-21), Department and Institute of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, Brazil

[17] Center for Interdisciplinary Research on Applied Neurosciences (NAPNA), University of São Paulo, São Paulo, Brazil

[18] MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK

[19] Department of Psychiatry, University of Münster, Münster, Germany

[20] Research Group in Psychiatry, Department of Psychiatry, Faculty of Medicine, Universidad de Antioquia, Medellín, Antioquia, Colombia

[21] Department of Psychiatry and Psychotherapy, Philipps-University Marburg, Marburg, Germany

[22] NeuroSpin, CEA, Paris-Saclay, Gif sur Yvette, France

[23] Department of Neurology, Oslo Universisty Hospital, Oslo, Norway

[24] Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA

[25] Desert-Pacific Mental Illness Research, Education, and Clinical Center, VA San Diego Healthcare System, La Jolla, CA, USA

[26] Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain

[27] Cardiff University Brain Research Imaging Centre, Cardiff University, Cardiff, UK

[28] Neuroscience Research Australia, Sydney, NSW, Australia

[29] School of Medical Sciences, University of New South Wales, Sydney, NSW, Australia

[30] Department of Psychiatry, Yale University, New Haven, CT, USA

[31] Olin Neuropsychiatric Research Center, Institute of Living, Hartford Hospital, Hartford, CT, USA

[32] Institut Pasteur, Unité Perception et Mémoire, Paris, France

[33] INSERM U955 Team 15 'Translational Psychiatry', University Paris East, APHP, CHU Mondor, Fondation FondaMental, Créteil, France

[34] Neuroscience Institute, University of Cape Town, Cape Town, South Africa

[35] Translational Neuroscience Group, Department of Psychiatry and Mental Health, Cape Town, South Africa

[36] School of Psychiatry, University of New South Wales, Sydney, NSW, Australia

[37] Department of Psychiatry and Behavioural Sciences, University of New Mexico, Albuquerque, NM, USA

[38] Mood Disorders Program, Hospital Universitario San Vicente Fundación, Medellín, Antioquia, Colombia

[39] Department of Psychiatry, University of Texas Health Science Center at Houston, Houston, TX, USA

[40] Psychosomatic Unit, Division of Mental Health and Dependence, Oslo University Hospital and University of Oslo, Oslo, Norway

[41] University of Oslo, Institute of Clinical Medicine, Oslo, Norway

[42] Black Dog Institute, Prince of Wales Hospital, Sydney, NSW, Australia

[43] Research Group, Instituto de Alta Tecnología Médica (IATM), Medellín, Antioquia, Colombia

[44] Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA

[45] Department of Psychiatry, SA MRC Unit on Risk & Resilience in Mental Disorders, University of Cape Town, Cape Town, South Africa

[46] Western Cape Department of Health, Valkenberg Hospital, Cape Town, Western Cape, South Africa

[47] Department of Psychiatry, University of Vermont, Burlington, VT, USA

[48] Department of Child and Adolescent Psychiatry/Psychology, Erasmus Medical Centre, Rotterdam, The Netherlands

[49] Department of Psychology, University of Oslo, Oslo, Norway

[50] Instituto de Ensino e Pesquisa, Hospital Sírio-Libanês, Sao Paulo, Brazil