Schwarz et al. *Translational Psychiatry* (2019)9:12
DOI 10.1038/s41398-018-0225-4

Translational Psychiatry

ARTICLE Open Access

Reproducible grey matter patterns index a multivariate, global alteration of brain structure in schizophrenia and bipolar disorder

Emanuel Schwarz[1], Nhat Trung Doan[2], Giulio Pergola[3], Lars T Westlye[2,4], Tobias Kaufmann[2], Thomas Wolfers[5,6], Ralph Brecheisen[7], Tiziana Quarto[3,8], Alex J Ing[9], Pasquale Di Carlo[3], Tiril P Gurholt[2], Robbert L Harms[10], Quentin Noirhomme[10], Torgeir Moberget[2], Ingrid Agartz[2,11,12], Ole A Andreassen[2], Marcella Bellani[13,14], Alessandro Bertolino[3,15], Giuseppe Blasi[3,16], Paolo Brambilla[17], Jan K Buitelaar[18,19], Simon Cervenka[11], Lena Flyckt[11], Sophia Frangou[20], Barbara Franke[18,21], Jeremy Hall[22], Dirk J Heslenfeld[23], Peter Kirsch[24,25], Andrew M McIntosh[26,27], Markus M Nöthen[28,29], Andreas Papassotiropoulos[30,31,32,33], Dominique J-F de Quervain[31,32,34], Marcella Rietschel[35], Gunter Schumann[9], Heike Tost[1], Stephanie H Witt[35], Mathias Zink[1,36] and Andreas Meyer-Lindenberg[1], The IMAGEMEND Consortium, Karolinska Schizophrenia Project (KaSP) Consortium

Abstract

Schizophrenia is a severe mental disorder characterized by numerous subtle changes in brain structure and function. Machine learning allows exploring the utility of combining structural and functional brain magnetic resonance imaging (MRI) measures for diagnostic application, but this approach has been hampered by sample size limitations and lack of differential diagnostic data. Here, we performed a multi-site machine learning analysis to explore brain structural patterns of T1 MRI data in 2668 individuals with schizophrenia, bipolar disorder or attention-deficit/hyperactivity disorder, and healthy controls. We found reproducible changes of structural parameters in schizophrenia that yielded a classification accuracy of up to 76% and provided discrimination from ADHD, through it lacked specificity against bipolar disorder. The observed changes largely indexed distributed grey matter alterations that could be represented through a combination of several global brain-structural parameters. This multi-site machine learning study identified a brain-structural signature that could reproducibly differentiate schizophrenia patients from controls, but lacked specificity against bipolar disorder. While this currently limits the clinical utility of the identified signature, the present study highlights that the underlying alterations index substantial global grey matter changes in psychotic disorders, reflecting the biological similarity of these conditions, and provide a roadmap for future exploration of brain structural alterations in psychiatric patients.

Correspondence: Emanuel Schwarz (emanuel.schwarz@zi-mannheim.de) or Andreas Meyer-Lindenberg (Andreas.Meyer-Lindenberg@zi-mannheim.de)
[1]Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany
[2]Norwegian Centre for Mental Disorders Research (NORMENT), KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Institute of Clinical Medicine, University of Oslo, Oslo, Norway
Full list of author information is available at the end of the article.

Introduction

Schizophrenia is a severe neuropsychiatric disorder affecting approximately 0.7% of the population[1]. A large spectrum of experimental approaches has been used to identify neural alterations in schizophrenia[2,3]. Among these, magnetic resonance imaging (MRI) has received

particularly strong interest[4] due to its non-invasiveness, high efficiency in acquiring brain-wide information on structure and function, and the ubiquitous availability of scanners, enabling the accumulation of large sample sizes. Meta-analyses of MRI data have demonstrated the presence of widespread brain-structural changes in patients[5–14], and machine learning, whereby combined effects of numerous predictors can be exploited, has been used to identify predictive patterns that explain a substantial amount of schizophrenia-associated variation[15,16].

With a few notable exceptions[17–19], pattern recognition studies on brain MRI data have only been performed in single-site studies that demonstrate substantial variability in accuracy of case-control classification between studies. A recent meta-analysis suggests that this variability may be attributable to small sample sizes, with larger studies converging at 70- 80% accuracy[15]. The latter accuracy is consistent with a recent, large-scale multi-site investigation showing reproducible brain-structural differences between individuals with schizophrenia and healthy controls[20]. These limitations in accuracy pose a significant challenge to translate psychiatric MRI tools for diagnostic and predictive applications into clinical practice. The clinical utility of such tools strongly depends on their value for everyday clinical decision making, which usually requires differential diagnosis among different disorders rather than control/case discriminations. Therefore testing diagnostic specificity is of paramount importance[21]. Bipolar disorder has particularly high differential diagnostic relevance for schizophrenia and previous studies have provided promising evidence that structural differences in schizophrenia show specificity against this disorder[22–24]. Furthermore, symptoms of attention-deficit/hyperactivity disorder (ADHD) are among the frequent precursors of schizophrenia[25–31] during adolescence, but have less differential diagnostic relevance in adult individuals. The three conditions show substantially shared genetic risk, and conjointly map to a spectrum of neuropsychiatric disorders with brain structure alterations associated with genetic and environmental risk factors[32].

Based on these considerations, the collaborative FP7 project IMAging GEnetics for MENtal Disoders (IMAGEMEND) has assembled a large, multimodal database that comprises neuroimaging data on cohorts of individuals with schizophrenia and bipolar disorder, adolescent as well as adult individuals with ADHD, and healthy controls[33]. The primary focus of the project is the identification of multivariate biological signatures that can aid diagnosis of these disorders. Using this resource, we analyzed structural MRI data from 2668 individuals in the present study.

Our primary aims were 1) to identify brain structural patterns that can reproducibly differentiate individuals with schizophrenia from controls, 2) explore their diagnostic specificity with regard to other disorders and 3) to identify the underlying brain structures driving successful classification. The availability of matched case-control data from several sites allowed application of a leave-site-out procedure, meaning that data from all but one site were iteratively used for algorithm training and the remaining data used for testing. This was aimed at the identification of differences robust against between-site variability. In order to make use of the complementary information provided by the different measures, we included both 1) FreeSurfer-based measures of cortical morphometry (cortical thickness, surface area and volume) and global and subcortical volumetry as provided by Freesurfer[34], and 2) voxel-based morphometry (VBM) as provided by Statistical Parametric Mapping (SPM)[35]. We also compared two machine learning strategies: (I) random forest machine learning, which captures non-linear and multiplicative effects of predictors and yields an efficient ranking of important predictors, and (II) support vector machines (SVM), the most commonly and successfully applied linear tool in machine learning studies on brain structure[36].

Materials and methods

Cohorts

This study comprised eight cohorts with a total of 2668 participants (consisting of patients with schizophrenia (n = 375, cases in cohorts I-IV), bipolar disorder (n = 222, part of cohort VIII), ADHD (n = 342, cases in cohorts V and VI), as well as healthy control subjects (n = 1729, cohorts I to VIII; n = 368 of these in cohorts I-IV) demographic details are shown in Supplementary Table 1; recruitment details are shown in Supplementary Table 2). All participants gave written, informed consent and the study received approval from the local ethics committees of the participating institutions.

Data pre-processing

Pre-processing of all T1-weighted images was performed centrally at the same site (University of Oslo, Norway) using FreeSurfer 5.3 (http://surfer.nmr.mgh.harvard.edu)[34]. All datasets underwent visual assessment and minor manual intervention to correct for segmentation errors wherever necessary. Data with significant low quality due to, e.g., motion artifacts and image distortions were excluded. Cortical parcellation was performed using the Desikan–Killiany atlas[37,38], and subcortical segmentation was performed based on a probabilistic atlas[39]. The mean thickness, sum surface area, and volume for each cortical region-of-interest (ROI), as well as the volume of subcortical structures were computed, resulting in a set of 152 FreeSurfer features (Supplementary Table 4).

An important question of the present study was whether signatures that combined the effects of multiple brain

structures could be represented through regionally non-specific, 'global grey-matter features'. For this, we manually selected 20 of such 'global features' and these are detailed in Supplementary Table 11. Additionally, the per-subject median of all ventricle features was used as readout for global ventricle size. Furthermore, for VBM- and FreeSurfer-based analyses we determined separately the per-subject median across all features, resulting in a 'median feature', resulting in a set of 22 'global features' in total. To avoid feature redundancy, bilateral features were removed if both uni-lateral features were available.

The dataset was also processed each using VBM[35] as implemented in the CAT12 toolbox (http://dbm.neuro.uni-jena.de/cat/), SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) and MATLAB 2014a (Mathworks, Sherborn, MA, USA) to derive the grey matter (GM) maps. As input, we used the *nu.mgz* volume, an intensity-normalized volume adjusted for the non-uniformity in the original T1-images, obtained from the FreeSurfer pre-processing pipeline (https://surfer.nmr.mgh.harvard.edu/fswiki/ReconAllOutputFiles). Briefly, this volume was tissue-segmented into GM, white matter (WM) and cerebrospinal fluid maps. The modulated GM maps were subsequently registered to the Dartel template, which is based on 550 healthy subjects from the IXI database (http://brain-development.org/ixi-dataset/), using affine registration followed by the Dartel non-rigid registration algorithm[40]. The mean GM density was then computed for each region-of-interest as defined in the Automated Anatomical Labeling (AAL) atlas[41], resulting in a set of 122 VBM features (Supplementary Table 3).

Matching, covariate adjustment and normalization

An overview of the pre-processing and machine learning pipeline is shown in Fig. 1. Cohorts I to IV were used for subsequent training of machine learning algorithms. In cohorts II to IV, propensity score matching (using the R library *MatchIt*[42]) was used to create schizophrenia-control datasets, 1:1 matched on age and sex. Matching was performed separately for each cohort. No matching was performed in cohort I, since it comprised fewer controls than patients and showed no significant case-control differences regarding age and sex. Controls not selected during the matching process were retained for validation of algorithms (cohort VIII).

Covariate adjustment was performed in two steps. The first step was aimed at removing the effects of covariates relevant within a given dataset. For this, linear regression was used to construct normalization models in the matched case-control data (Supplementary Figure 1). Each feature was regressed against age, age^2, sex, and total intracranial volume (ICV, derived from FreeSurfer; this covariate was not included for thickness features derived from FreeSurfer processing). Normalization models were built separately for the cohorts used for training (i.e. during the leave-site-out procedure described below as well as for prediction of the schizophrenia classifier into the validation cohorts), and the resulting coefficients were averaged to obtain a final model per brain feature. These models were then applied to residualize the features in the training as well as the test data. Subsequently, ICV was added as a feature to the residualized training and test data. In the second covariate adjustment step, the effects of between-dataset variables (field strength and scanner vendor) were removed. Using data from the previous step as input, linear models were built to residualize all training data and adjust the test data accordingly. During the leave-site-out testing procedures, as well as for testing classifiers in validation data, the test data were not used to generate normalization models and remained independent. The objective of this two-step procedure was to appropriately account for the effect of potential confounders, without using site-information as additional covariate. This is essential for potential clinical application of a diagnostic tool, when subjects from sites are tested that are not part of the training data. In this case, adjustment against a site-covariate cannot be performed. In a secondary analysis, we set the means of each feature in a given test dataset artificially to 0 (for training data this is already fulfilled due to the residualization procedure). With this we tested whether not using test data for

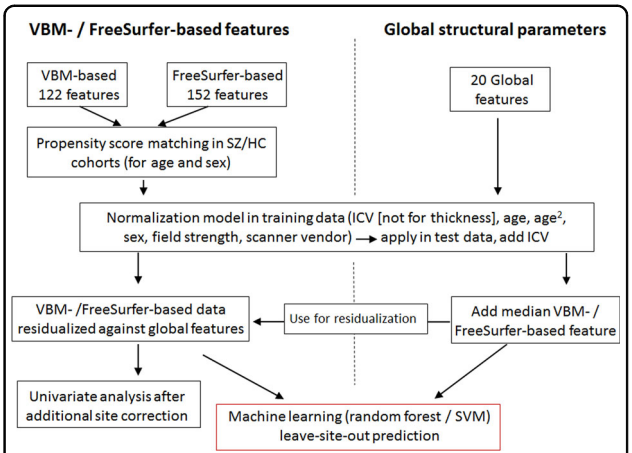

Fig. 1 Overview of analysis procedure. Subjects were first propensity score matched and VBM- / FreeSurfer-based features were then normalized against potential confounders. Normalization models were built in training data only and these models were subsequently applied to adjust the test data. The same normalization strategy was applied for global structural parameters, which were subsequently used to remove the global structural signal from VBM- / FreeSurfer-based features. The resulting data was used for leave-site-out cross-validation analyses. For univariate analyses, as well as for machine learning analyses performed on the entire dataset, data were additionally corrected for a site factor, to account for the impact of site differences (see methods)

building of normalization models impacted on classification performance.

For the machine learning analyses performed on the entire, matched dataset (i.e. for out-of-bag performance evaluation, where accuracy estimates were obtained from observations not selected during the repeated bootstrapping part of the random forest classification procedure, see below), we excluded the impact of a site factor through residualization using linear models, in addition to the covariate adjustment described above. For this residualization, site and scanner vendor were both included as covariates. Such corrected data was also used for the univariate analyses (see below). For principal components analysis, which was applied to explore the global similarity between VBM- and FreeSurfer-based features, data were additionally normalized against diagnosis and subsequently standardized.

Univariate analysis

Univariate analyses were performed to assess the extent of change in individual brain-structural measures prior to and following adjustment for global structural parameters. Univariate analysis was performed on data residualized as described above, to increase comparability against the features' importance determined by machine learning. Case-control differences were evaluated using Student's t-tests and *P*-values were adjusted for the False Discovery Rate (FDR) according to the method of Benjamini and Hochberg[43]. The adjustment was performed separately for VBM- and FreeSurfer-based features.

For the univariate analysis of the features following removal of the global structural signal, we first corrected the global structural features using the same steps described above. These corrected global structural features were then used to adjust the VBM- and FreeSurfer-based features, and the resulting residuals were used for the univariate analysis.

Machine learning – cross-validation and accuracy estimation

Several different procedures were employed to train and test machine learning algorithms: a) 'within-site' classification, where algorithms were trained and tested separately in each given cohort (using cohorts I-IV for schizophrenia-control classification, cohort VIII (selecting University of Oslo data only) for bipolar disorder-control classification, and cohorts V and VI for ADHD-control classification). b) 'Leave-site-out' classification in cohorts I-IV. c) Prediction of a schizophrenia-control classifier in independent test data (the classifier was trained in cohorts I-IV and tested in cohorts V-VIII).

For procedures a) and b), performance of machine learning algorithms was assessed by comparing the predicted class membership against the real class-membership. For 'within-site' classification, this was performed using bootstrapping.

The Receiver Operating Characteristic Area Under Curve (AUC) was determined to quantify accuracy (using the R library *pROC*[44]). For leave-site-out classification, we additionally determined the mean of sensitivity and specificity to explore whether predicted class probabilities were shifted across cohorts.

For procedure c), accuracy was determined as the specificity, i.e. the percentage of subjects correctly classified as being not affected by schizophrenia.

Machine learning – random forests

Random forest is a machine learning tool suitable for classification and regression[45]. It combines the output of a large number of individual classification/regression trees, each of which are built on randomly selected subsets of observations and predictors. The random forest can naturally incorporate interactions between predictors, allows efficient ranking of predictor importance and has been shown to be one of the most accurate classification tools on a large variety of data sets[36].

Random forest machine learning (using the R package *randomForest*[46]) was performed in a site-stratified manner using 5000 trees and the default value for the *mtry* parameter (no tuning of random forest parameters was performed). The number of trees was chosen based on the observation that larger tree numbers do not significantly improve performance[47]. Site-stratification was performed such that for building each tree, an equal number of subjects (equal to the sample size of the smallest training cohort) were randomly drawn without replacement from the data of each site. We determined the importance of the features for prediction during this procedure using the Gini index, a measure of how much a given feature impacts the correct class separation, when used for a split during the tree-building process[48]. Selection of the most important predictors was performed using the R package *varSelRF*[49], also using 5000 trees, and default settings otherwise. During this procedure, the least important variables are successively removed from the model. The optimal number of variables is chosen for the solution where the out-of-bag error is equal to the lowest observed error rate, plus one standard deviation. This leads to a solution with close to optimal error rate but with a lower number of predictors, a scenario generally thought to be beneficial for the generalizability of the classifier. The Gini-index-derived variable importance measure was also used to assess the similarity of features selected by within-site classification. For this, we determined the median Pearson correlation of the variable importance measures across cohorts.

To explore the diagnostic specificity of important variables, we first selected the top *m* (with m being

determined via random forest variable selection; $m = 14$ for VBM-based and m = 11 for FreeSurfer-based features, respectively) variables from the schizophrenia-control comparison. We then determined the Wilcoxon rank sum statistic comparing the importance of these variables against the remaining variables in bipolar disorder, adolescent as well as adult ADHD. To test significance, a 5,000-fold permutation of diagnostic labels was performed. During each repetition, variable importance was re-calculated for the three non-schizophrenia case-control comparisons and the determination of rank sum statistics was repeated. Empirical *P*-values were then calculated as the frequency of permutation rank sum statistic at least as high as those determined from non-permuted data.

Random forest regression was used to determine the amount of variance that could be predicted in individual VBM- and FreeSurfer-based features using the global structural parameters. The explained variance was determined from out-of-bag predictions. For this analysis, the same covariate-adjusted data were used as for the univariate analysis (see above). Accordingly, the global structural parameters were also additionally residualized against a site factor.

Machine learning – Support Vector Machines

A support vector machine is a classification tool that aims to identify a decision boundary with maximal margin between the boundary and observations from a given class[50]. The boundary is defined based on the most proximal observations, making classification insensitive to data variations or outliers, resulting in frequently superior generalization performance[36]. Linear SVM is relatively robust to overfitting and was, in the present study (using the R package *e1071*[51]), tuned using 10-fold cross-validation to optimize the cost parameter (choosing among values from the log sequence between 10^{-5} and 10^{5}). Parameter optimization was performed in training data only.

Exploring the impact of global structural parameters on classification

To explore the effect of the 22 global structural features on classification, these features were adjusted for confounding variables using the same procedure applied for VBM- and FreeSurfer-based features (i.e. residualization against age, age^2, sex, gender, ICV, field strength, and scanner vendor). VBM- and FreeSurfer-based features were subsequently residualized against the covariate-adjusted global features using additive linear models. To explore the impact of this residualization procedure *per se*, it was repeated 1000 times with row order-permuted global features. Similarly, to explore the significance of the accuracy obtained after residualization, the procedure was repeated 1000 times with permuted diagnostic labels. Finally, to explore the classification accuracy obtained from global-features only, we applied random forest machine learning (as described above) using the covariate-adjusted global features.

Results

Brain structural neuroimaging data from a total of 2668 subjects were analyzed. Sample details are presented in Supplementary Tables 1 and 2. The data were processed to extract either 122 VBM-based or 152 FreeSurfer-based morphometry features (Fig. 1, Supplementary Tables 3 and 4, ICV was added as a predictor to each feature set). Machine learning was used to identify structural patterns that could be used to differentiate individuals with schizophrenia from controls and to establish the diagnostic specificity against bipolar disorder and ADHD.

Case-control differences, schizophrenia classification and diagnostic specificity *Univariate case-control differences*

the univariate analysis of matched cases and controls from cohorts I to IV demonstrated significant alterations in VBM-based features of individuals with schizophrenia (Supplementary Tables 3 and 4). A total of 110 of the 123 features showed significant alteration at FDR < 0.05. Similarly, for FreeSurfer-based features, 105 of the 153 features were significant at this threshold.

Machine-learning classification

Using random forest machine learning, we first performed a within-site classification of participants with schizophrenia and controls and found AUC values obtained from out-of-bag predictions ranging from 0.58 to 0.82 for VBM-based and from 0.58 to 0.80 for FreeSurfer-based features, respectively (Supplementary Table 5). Permutation analysis showed that accuracy estimates were significant for three of the four cohorts (Supplementary Table 5). When all case-control cohorts were combined into a single dataset, the AUC obtained from out-of-bag predictions was 0.73 ($P < 0.001$) for VBM-based and 0.72 ($P < 0.001$) for FreeSurfer-based morphometry, respectively. When VBM- and FreeSurfer-based features were combined into a single dataset, the resulting AUC was 0.74 ($P < 0.001$). We further found that features were more consistently selected as important predictors for VBM data (median correlation of variable importance measures across the four cohorts of 0.11) compared to FreeSurfer data (mean correlation -0.02).

Leave-site-out classification

We tested the classification accuracy when all but one of the case-control datasets were used for training. This leave-site-out cross-validation yielded median AUC

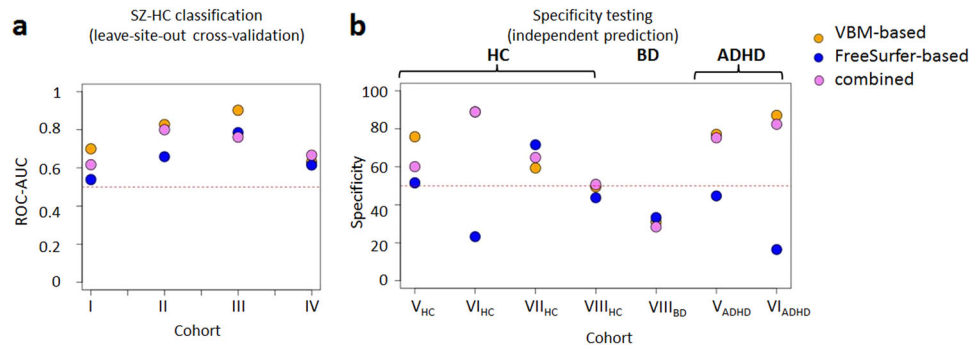

Fig. 2 Accuracy of schizophrenia classifier using VBM- and FreeSurfer-based morphometry features. a) Leave-site-out cross-validation performance measured as the ROC-AUC. **b** Specificity of schizophrenia-control classifier (trained on all SZ-HC cohorts) for prediction in independent cohorts. The red horizontal line demonstrates 50% ROC-AUC or specificity, respectively. The classification was based on random forest machine learning. SZ: schizophrenia; BD: bipolar disorder; ADHD: attention-deficit/ hyperactivity disorder; HC: healthy controls

estimates of 0.76 (range 0.63 to 0.90) and 0.64 (range 0.54 to 0.78) for VBM- and FreeSurfer-based morphometry features, respectively. The median AUC for the combined feature set was 0.71 (range 0.62 to 0.80) (Fig. 2a). For VBM-based data, the observed accuracy corresponded to a sensitivity-specificity mean with a median of 0.70 across cohorts I-IV. We observed that sensitivity and specificity varied substantially across cohorts (Supplementary Table 6). In FreeSurfer-based data, this was even more pronounced with a corresponding estimate of 0.52, showing that the optimal cut-off for classification differed across cohorts (Supplementary Figure 2). This was likely due to shifts of structural volume means across cohorts. The normalization models aim to set structure mean values in the test data to zero, but this is not guaranteed as test data were not used for building the normalization models. Setting test data means to zero (a strategy commonly employed in machine learning) resolved the sensitivity-specificity imbalance (sensitivity-specificity mean with a median of 0.76, 0.71 and 0.71 for VBM-, FreeSurfer and combined data, respectively. AUC values were 0.79, 0.75 and 0.78, respectively; see Supplementary Table 7).

Specificity testing in independent test cohorts

For VBM-based features, the application of an algorithm trained on all four training cohorts resulted in accuracies ranging from 50% to 89% (median 68%) in four independent cohorts of healthy controls (Fig. 2b, Supplementary Table 8). The algorithm showed limited specificity against bipolar disorder as 69% of the 222 individuals were assigned to the schizophrenia class. To explore potential associations between prediction accuracy and the presence of psychotic features among individuals with bipolar disorder, we identified subsets of individuals with severe psychosis (n = 28) and individuals without psychotic features (n = 48). However, we found no evidence that accuracy significantly differed between these clinical groups ($P = 0.63$).

In contrast, when applying the algorithm to adult (n = 85) and adolescent (n = 257) subjects with ADHD, schizophrenia classification showed similar accuracy (87% and 77% correctly classified as not belonging to the schizophrenia class) as for healthy control subjects. Notably, classification based on FreeSurfer-based morphometry features showed substantially poorer accuracy in most independent validation cohorts (Fig. 2b, Supplementary Table 8). As for leave-site-out classification, this was due to mean shifts of covariate-adjusted data that affected FreeSurfer-based morphometry features important for schizophrenia classification and is exemplified for amygdala volumes in Supplementary Figure 3.

Comparison between classifier types

To explore whether prediction results were influenced by the choice of the algorithm, we replaced the site-stratified random forest with a non-site-stratified, linear SVM. This showed that across all conducted tests, SVM outperformed random forest classification by a small margin (Supplementary Table 6, Supplementary Figure 4). Notably, linear SVM application also showed an improved specificity of the schizophrenia classification against bipolar disorder (specificity between 48 and 55%, Supplementary Table 6, Supplementary Figure 4).

Case-control classification of differential diagnoses

VBM-based data showed limited utility for a meaningful differentiation of bipolar disorder (AUC of 0.63, derived from random forest out-of-bag prediction), adult (AUC = 0.58), or adolescent (AUC = 0.62) ADHD from healthy controls within the respective, propensity score-matched cohorts. On the same cohorts, similar performance estimates (AUC of 0.66, 0.56, and 0.63 respectively) were obtained for FreeSurfer-based features.

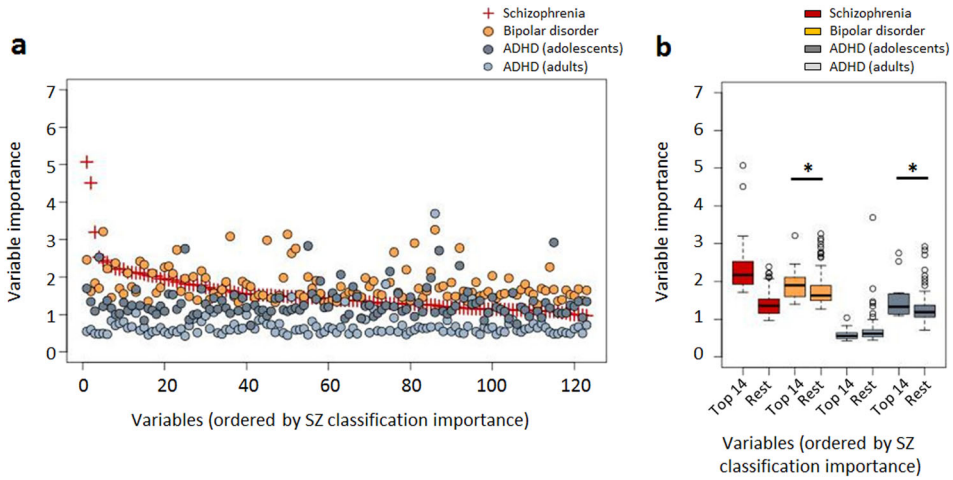

Fig. 3 VBM-based variable importance for classification. a Random-forest variable importance for the schizophrenia vs. control (red, used to order the x-axis), the bipolar disorder vs control and the ADHD vs control comparisons. **b** Boxplot of random-forest variable importance measures, comparing the 14 most important schizophrenia predictors against the remaining predictors in bipolar disorder and ADHD. The asterisk indicates significance determined from permutation testing. Since variable importance was determined from the schizophrenia-control comparison, no significance estimate is shown for the corresponding boxplot

Exploration of features important for classification

The random forest variable importance derived from the site-stratified classifiers based on all case-control cohorts was used to identify the features most relevant for classification. The ranked variable importance measures derived from VBM-based morphometry data are shown in Fig. 3a (and Supplementary Table 9). Using random forest feature selection, we found 14 VBM-based features (11 for FreeSurfer-based data) to be of particular importance for classification, i.e. the respectively smallest feature sets leading to the minimum error rate plus one standard deviation (see methods). Figure 3a further displays the importance of VBM-based features for classification of bipolar disorder (propensity score-matched patients and controls from University of Oslo bipolar disorder and control data part of cohort VIII, n = 444) and ADHD (propensity score-matched patients and controls from cohorts V (adolescent subjects), n = 322, and VI (adult subjects), n = 170). The top 14 features for schizophrenia-control classification had also significantly higher importance for bipolar disorder-control as well as the adolescent subjects with ADHD vs. controls classification ($P = 0.011$ and $P = 0.008$, respectively; permutation test, Fig. 3b), compared to the remaining features. In contrast, these features were of no significant importance for the adult ADHD-control classification ($P = 0.857$, Fig. 3b). Supplementary Figure 5 displays the variable importance measures derived from FreeSurfer-based morphometry data (Supplementary Table 10), showing a similar pattern for schizophrenia markers and those for bipolar disorder ($P = 0.003$) as well as adult ($P = 0.196$) ADHD compared to VBM-based analysis. Notably, for FreeSurfer-based morphometry data, no overlap with adolescent ADHD markers was found ($P = 0.350$).

Relation between VBM-based and FreeSurfer-based predictors

Between the top-14 VBM-based and the top-11 FreeSurfer-based predictors for the schizophrenia-control classification, we found significant pairwise correlations (median Pearson's correlation coefficient of 0.16, using subjects from cohorts I to IV, after additional residualization against diagnosis). Accordingly, in this confounder-corrected dataset, the first principal components (PCs) of the top features (explaining 42% and 38% of variance in FreeSurfer-based and VBM-based features, respectively), were strongly correlated ($\rho = 0.43$, $P = 5.4 \cdot 10^{-34}$). This raised the question whether the numerous, individually weak structural predictors were related to a common global measure of brain structure. To explore this, we tested associations between the principal components and 22 global measures of brain structure and found highly significant correlations with the large majority of these measures (Fig. 4a, Supplementary Table 11). This effect was not due to residual confounding of any PC by total intracranial volume, age, age^2, sex, scanner vendor, field strength or recruitment site (all uncorrected $P > 0.12$).

Effect of global structural parameters on classification and univariate differences

We then explored, whether these global measures explained part of the multivariate signal that allowed case-control differentiation between patients and controls. Figure 4b shows that residualization of VBM- and

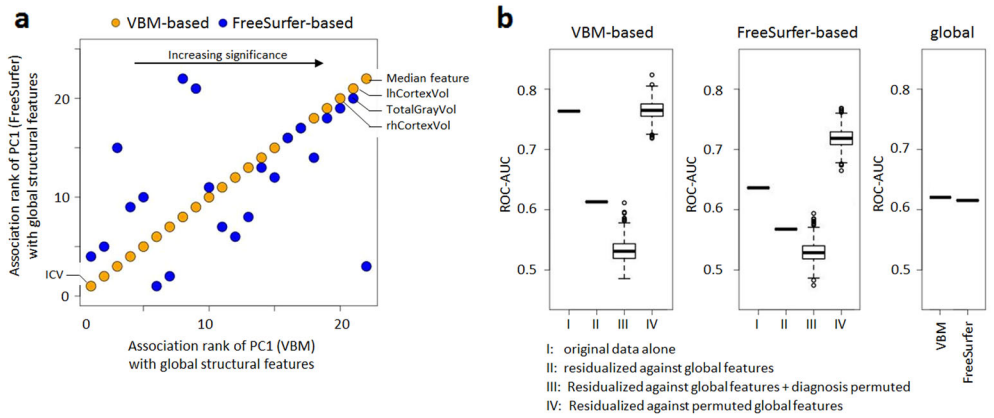

Fig. 4 Effect of global structural covariates on classification. a Comparison of associations between global structural features and the first principal components determined from the 14 selected VBM-based (orange; used to order the x-axis) and the 11 selected FreeSurfer-based (blue) features (see also Supplementary Table 1,0). **b** Effect of residualization against global structural features on classification performance and classification performance obtained from global features only. Notably, AUC values obtained from analyses with permuted diagnoses showed mean values > 0.5, which was due to chance associations in the comparatively small datasets. Furthermore, surface based features showed an increase in performance after residualization against permuted global features. This suggests features with poor cross-site reproducibility were coincidentally prioritized for classification in the original data and this was remedied in the residualized data. The two sets of global features were identical except for the addition of either a median VBM- or FreeSurfer-based feature

FreeSurfer-based features against the 22 global measures led to a decrease in classification performance (measured as the leave-site-out AUC determined on cohorts I to IV) from 0.76 to 0.61 (VBM-based) and from 0.64 to 0.57 (FreeSurfer-based), respectively. These AUC values were close to (VBM-based) or within (FreeSurfer-based) the range of those obtained after randomly permuting diagnostic grouping (Fig. 4b). Accuracy did not decrease substantially, when residualization was performed with permuted global covariates, showing that residualization against large covariate numbers did not per se have a substantial impact (Fig. 4b). Classification using covariate-corrected global features alone led to a leave-site-out AUC of 0.62, regardless of whether the median VBM- or the median FreeSurfer-based feature was included (Fig. 4b). This raises the question why global structural features were strong co-variates of case-control associations, but relatively poor predictors of diagnostic status when used alone. This effect was likely due to site-to-site variability of the global structural features, since random forest learning applied on the entire dataset yielded out-of-bag AUC values of 0.71 for both global structural parameter sets. These values were comparable to the out-of-bag estimates derived from similarly corrected VBM- (AUC = 0.73) or FreeSurfer-based (AUC = 0.72) features. This further supports the extent of signal shared between global features and individual brain structures.

Notably, the residualization against global features also led to substantial decrease in univariate significance (Supplementary Table 3). For VBM-based features, after residualization, FDR-corrected significance was only observed for a bilateral increase in the pallidum (left: $P_{FDR} = 2.5{\cdot}10^{-5}$; right: $P_{FDR} = 1.5{\cdot}10^{-4}$) and a decrease in the right hippocampus ($P_{FDR} = 0.026$). For FreeSurfer-based features, after residualization against global parameters, no significance was observed.

Prediction of individual structural features through global structural parameters

We explored whether individual brain structural features could be accurately predicted based on global structural parameters. Based on random forest regression, the global features explained a mean of 29% ± 13 (range 2.5% – 61.2%) of variance in VBM-based features and a mean of 29% ± 15 (range 0.0% – 64.8%) of variance in FreeSurfer-based features, respectively (Supplementary Tables 3 and 4). In VBM-based data, the variance explained by global features was further correlated with the mean size of the respective structure ($\rho_{VBM} = 0.33$; $P_{VBM} = 0.0002$; $\rho_{surface} = -0.06$; $P_{surface} = 0.44$; Spearman correlation, to prevent overdue influence of larger structures).

Discussion

The primary findings of this multi-site investigation were 1) the presence of reproducible brain-structural patterns that could differentiate individuals with schizophrenia from healthy controls, 2) the specificity of the patterns when applied on data from individuals with ADHD, and the lack thereof in bipolar disorder, 3) the significant overlap of markers important for classification of schizophrenia, bipolar disorder and adolescent ADHD

and 4) the finding that brain-structural changes were strongly associated with global structural parameters.

Based on brain-structural patterns, individuals with schizophrenia could be reproducibly differentiated from healthy controls, with a median AUC of up to 0.76. Performance estimates were derived from unbiased leave-site-out cross-validation and no test set data were used to determine parameters of covariate adjustment or machine learning models. Therefore, the obtained estimates are likely to reflect the performance of the algorithms, when tested in independent data. We observed that when test data were not used during generation of normalization models, sensitivity and specificity fluctuated substantially, which could be resolved by scaling of the test data. This, however, would require at least some data from a given test site to be available prior to testing algorithms in data from that site[20]. It should also be noted that biological heterogeneity resulting from the current diagnostic system limits the accuracy biological predictions can achieve, when aiming to reproduce clinical classifications, constituting a general caveat for the field.

The brain-structural patterns associated with schizophrenia showed significant lack of specificity against bipolar disorder, consistent with the substantial genetic and clinical overlap of the two disorders[30,31,52]. Notably, the signatures were specific against adolescent and adult ADHD. Subjects with ADHD, did not, however, show brain-structural alterations that could be used for accurate classification, nor did those with bipolar disorder. Despite this, the VBM-based feature sets most useful for classification of adolescent ADHD and schizophrenia showed significant overlap. Given the high specificity of the schizophrenia classifier against adolescent ADHD, this supports divergent profiles in the same feature set. A particular strength of the present study was that conclusions regarding differential diagnostic specificity against bipolar disorder were not confounded by site variability. Considering the observed specificity fluctuations during leave-site-out testing, it should, however, be noted that the preferential classification of subjects with ADHD as controls could have been influenced by between-site effects. Similarly, non-specificity of the schizophrenia classifier against bipolar disorder was determined in one cohort and requires further replication. Also, the lack of adolescent subjects in the training data may have confounded the accuracy observed in adolescent ADHD subjects.

We aimed to identify brain-structural features driving reproducible schizophrenia-control classification and to compare these between two different pre-processing strategies. We observed that these strategies led to identification of differential structural patterns but found that these alterations were, to a large extent, capturing overlapping global brain-structural alterations. Removing variation explained by measures of global structural properties also removed most of the identified multivariate signals. Notably, global structural parameters were strong confounders of VBM- and FreeSurfer-based feature associations, but were on their own relatively poor predictors of diagnosis. Our results indicate that this was, to a significant extent, due to between-site variability affecting the global signal. This effect may be due to the fact that the global signal combines multiple signals that are individually affected by site-specific effects (such as the shifts in mean measurement observed in the present study), creating an aggregate signal reflecting site idiosyncrasies. This, in turn, raises the important question to what extent global variables reflect the underlying biology vs. measurement factors (i.e. the signal to noise ratio) in structural imaging data. The observed case-control classification performance is consistent with previous large-scale analyses[15,20], thus it is unlikely that measurement uncertainty specific to the present study accounts for the global effects detected. Furthermore, GM differences have been observed in numerous studies investigating first-episode schizophrenia patients, suggesting that these effects are not primarily related to the specific clinical characteristics of the samples we examined [e.g[53–55]]. One possible interpretation of these results is that schizophrenia entails a combination of isometric and allometric structural changes which may vary between individuals and within patients across different stages of the illness. This explanation may account for the low effect sizes and effect heterogeneities of structural differences previously observed in schizophrenia. Another interpretation is that a shared biological component affecting global variables across multiple disorders discriminates controls from cases, but does not differentiate patients with different diagnoses. Accordingly, previous reports highlighted shared genetic components across multiple psychiatric disorders and personality traits[56,57]. In contrast, the present results may also be interpreted from the perspective of cross-cohort reproducibility. That is, the reduction in classifier accuracy through consideration of global structural features primarily relates to effects on reproducible alterations in GM features. Changes in individual sites, in contrast, may have persisted despite the normalization against the global signals. This interpretation raises the question whether this and previous studies had sufficient resolution, in view of the large site to site differences, to investigate reproducible regional effects. An improved imaging resolution could also allow identifying patterns of structural differences that show higher specificity between schizophrenia and bipolar disorder. A corollary of this view is the question whether, even assuming that structural imaging resolution yields sufficient signal to noise ratio to study regional effects, the correlations between regional and global variables caused by common

underlying biology and by shared measurement uncertainties can be meaningfully disentangled. For example, we found that identification of univariate changes was strongly dependent on global structural alterations. Importantly, if the global signal was indeed more affected by site specific experimental effects than individual brain structures, it would be challenging for single-site investigations or univariate statistics to appropriately account for this effect, limiting the possibility to reproduce findings across studies.

In this context, a limitation of the present study is the lacking incorporation of other data modalities, such as demographic, clinical or psycho-behavioral features, which could potentially have informed on the presence of patient subgroups or illness-dimensions in relation to brain-structural alterations. Similarly, future studies should explore the effects of antipsychotic treatment on GM, which have been observed in schizophrenia (i.e. ref. [9]) and are supported by data from animal models[58,59], but which have also been found in antipsychotic-native subjects[9]. An acerbation of disorder-intrinsic structural changes by medication may be a possible explanation why removal of the global signal almost completely removed structural differences. While this study explores the impact of different pre-processing strategies on machine learning analysis of brain-structural differences, it does not offer a comprehensive analysis of the broad spectrum of preprocessing methods currently available. The sensitivity of machine learning to the choice of preprocessing may contribute to the variability of such analyses as reported in previous studies. Another limitation of the present study is the fact that it involved already diagnosed patients. One of the most significant aspects of clinical utility will be the ability to accurately predict the transition from early signs to full-blown illness, such that appropriate treatment can be started earlier.

Finally, an interesting finding was that linear SVM application showed marginally better classification performance compared to RF machine learning. This suggests that classification did not profit from RF's ability to model complex interactions. Interestingly, schizophrenia classification using linear SVM also showed an improved specificity against bipolar disorder, which requires further validation in independent cohorts.

In conclusion, this study identified reproducible GM patterns that index a multivariate, global alteration of brain structure in schizophrenia and bipolar disorder, but are different from those seen in ADHD. These results may reflect the biological heterogeneity of schizophrenia and are consistent with previous observations of shared genetic determinants between these disorders. The results further demonstrate the need for appropriately accounting for the global signal during analysis of individual brain structures. They underline the importance of biologically dissecting these illnesses as a basis to redefine diagnostic boundaries using biological parameters. These efforts may benefit from integrative analyses of other relevant data modalities, including genetic risk measures or functional neuroimaging, which may yield more accurate and specific classifiers that have clinical utility. Also, substantial differences in the ability to derive reproducible brain-structural signatures were found when using VBM or FreeSurfer features derived from the same individuals, highlighting the importance of preprocessing strategies for machine learning analysis of brain-structural data. Finally, the present results highlight the need for a more in-depth analysis of how individual brain structures contribute to the pathophysiology of these psychiatric disorders.

Code availability

Code used for the analyses described in this manuscriptis available from the corresponding author upon request.

Acknowledgements

We thank all the patients and healthy volunteers for their willingness to participate in the study. We also wish to express our appreciation to the KaSP research nurses. We would further like to thank Dr. Axel Schaefer and Marina Cariello for their assistance with this study. This study was supported by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 602450 (IMAGEMEND, IMAging GENetics for MENtal Disorders) and the Deutsche Forschungsgemeinschaft (DFG), SCHW 1768/1-1. A.M.-L. was supported by the Deutsche Forschungsgemeinschaft (DFG) (Collaborative Research Center SFB 636, subproject B7); the German Federal Ministry of Education and Research (BMBF) through the Integrated Network IntegraMent (Integrated Understanding of Causes and Mechanisms in Mental Disorders) under the auspices of the e:Med Programme (BMBF Grant 01ZX1314A and 01ZX1314G); and the Innovative Medicines Initiative Joint Undertaking (IMI) under Grant Agreements no 115300 (European Autism Interventions—A Multicentre Study for Developing New Medications) and no 602805 (European Union-Aggressotype). This study made use of the Dutch sample of the International Multicentre persistent ADHD CollaboraTion (IMpACT). IMpACT unites major research centres working on the genetics of ADHD persistence across the lifespan and has participants in the Netherlands, Germany, Spain, Norway, the United Kingdom, the United States, Brazil, and Sweden. The Dutch IMpACT node is supported by grants from the Netherlands Organisation for Scientific Research (NWO; grants 433-09-229 and 016-130-669 to BF), from the European Community's Seventh Framework Programme (FP7/2007-2013) (grant agreements no 278948 (TACTICS), no 602450 (IMAGEMEND), and no 602805 (Aggressotype)) and Horizon 2020 Programme (grant agreements no 643051 (MiND) and no 667302 (CoCA)). This research also receives funding from the European College of Neuropsychopharmacology (ECNP) Network 'ADHD across the Lifespan' and the National Institutes of Health (NIH) Consortium grant U54 EB020403, supported by a cross-NIH alliance that funds Big Data to Knowledge Centers of Excellence. The NeuroIMAGE study, also contributing data to this study, represents the longitudinal follow-up of the Dutch subsample of the International Multicentre ADHD Genetics (IMAGE) project. PIs of NeuroIMAGE are Jan Buitelaar and Barbara Franke (Radboud University Medical Center, Nijmegen), Jaap Oosterlaan and Dirk Heslenfeld (Vrije Universiteit Medical Centre, Amsterdam), and Pieter Hoekstra and Catharina Hartman (University Medical Centre Groningen). NeuroIMAGE is supported by grants from The Netherlands Organization for Health Research and Development (ZonMw 60-60600-97-193), the Netherlands Organization for Scientific Research (NWO, grants 1750102007010, 433-09-242 and 056-13-015), and by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 278948 (TACTICS), 602450 (IMAGEMEND), 602805 (AGGRESSOTYPE), 603016 (MATRICS), and Horizon 2020

(grant agreement 643051 (MiND) and 642996 (BRAINVIEW) research programmes. T.P.G. acknowledges funding from The Research Council of Norway (grant #223273) and the KG Jebsen Foundation. J.O. acknowledges funding by NIH Grant R01MH62873, NWO Large Investment Grant 1750102007010 and an NWO Brain & Cognition grant (056-24-011), the European Union 7th Framework programs AGGRESSOTYPE (602805) and MATRICS (603016), and by grants from Radboud University Medical Center, University Medical Center Groningen and Accare, and Vrije Universiteit Amsterdam. L.F. acknowledges funding by Söderbergs Königska Stiftelse, Stockholm County Council (ALF, PPG). H.F.B. acknowledges funding by Söderbergs Königska Stiftelse, Centre for Psychiatry Research (post doc stipendium). S.C. acknowledges funding by The Swedish Research Council (523-2014-3467) and the Stockholm County Council (20160328). P.K. acknowledges funding by the DFG (KI 576/14-2). T.K. acknowledges funding by the Research Council of Norway (grants #213837 and #223273 to PI Ole Andreassen). J.H. acknowledges funding by the Wellcome Trust as well as the MRC. D.J.F.d.Q acknowledges funding by the Swiss National Science Foundation. G.P. acknowledges funding by Fondazione CON IL SUD, and Hoffmann-La Roche. PB was partially supported by grants from the Italian Ministry of Health (RF-2011-02352308). P.M.T. acknowledges funding by NIH grant U54 EB020403. F.D. acknowledges funding by the German Federal Ministry of Education and Research (BMBF) grant 01ZX1314A/01ZX1614A. A.R. acknowledges funding by the "Capitale Umano ad Alta Qualificazione" grant awarded by Fondazione Con Il Sud. E.G.J. acknowledges funding by the Swedish Research Council, a regional agreement on medical training and clinical research between Stockholm County Council and Karolinska Institutet, and the HUBIN project. The HUBIN and KaSP studies were supported by the Swedish Research Council.

Author details

[1]Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany. [2]Norwegian Centre for Mental Disorders Research (NORMENT), KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [3]Department of Basic Medical Sciences, Neuroscience and Sense Organs, University of Bari Aldo Moro, Bari, Italy. [4]Department of Psychology, University of Oslo, Oslo, Norway. [5]Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands. [6]Donders Center for Cognitive Neuroimaging, Radboud University, Nijmegen, The Netherlands. [7]Maastricht University Medical Center, Maastricht, The Netherlands. [8]Cognitive Brain Research Unit, Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [9]Centre for Population Neuroscience and Stratified Medicine (PONS) and MRC-SGDP Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. [10]Brain Innovation B.V., Maastricht, The Netherlands. [11]Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, & Stockholm County Council, Stockholm, Sweden. [12]Department of Psychiatry Research, Diakonhjemmet Hospital, Oslo, Norway. [13]Section of Psychiatry, Azienda Ospedaliera Universitaria Integrata Verona, Verona, VR, Italy. [14]Department of Neurosciences, Biomedicine and Movements Sciences, University of Verona, Verona, VR, Italy. [15]Institute of Psichiatry, Policlinico Bari, Azienda Ospedaliero Universitaria Consorziale Policlinico Bari, Bari, BA, Italy. [16]Azienda Ospedaliero-Universitaria Consorziale Policlinico, Bari, Italy. [17]Department of Neurosciences and Mental Health, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, University of Milan, Milan, Italy. [18]Donders Institute for Brain, Cognition and Behaviour, Radboudumc, Nijmegen, The Netherlands. [19]Karakter Child and Adolescent Psychiatry University Center, Nijmegen, The Netherlands. [20]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [21]Departments of Human Genetics and Psychiatry, Radboud University Medical Center, Nijmegen, The Netherlands. [22]Neuroscience and Mental Health Research Institute, Cardiff University, Maindy Road, Cardiff CF24 4HQ, UK. [23]Department of Cognitive Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [24]Department of Clinical Psychology, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Heidelberg, Germany. [25]Bernstein Center for Computational Neuroscience Heidelberg-Mannheim, Mannheim, Germany. [26]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh EH10 5HF, UK. [27]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, George Square, Edinburgh EH8 9JZ, UK. [28]Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany. [29]Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany. [30]Division of Molecular Neuroscience, Department of Psychology, University of Basel, CH-4055 Basel, Switzerland. [31]Transfaculty Research Platform Molecular and Cognitive Neuroscience, University of Basel, Basel, Switzerland. [32]Psychiatric University Clinics, University of Basel, CH-4055 Basel, Switzerland. [33]Department Biozentrum, Life Sciences Training Facility, University of Basel, CH-4056 Basel, Switzerland. [34]Division of Cognitive Neuroscience, Department of Psychology, University of Basel, CH-4055 Basel, Switzerland. [35]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Heidelberg, Germany. [36]District Hospital Mittelfranken, Department of Psychiatry, Psychotherapy and Psychosomatics, Ansbach, Germany

The IMAGEMEND Consortium

Francesco Bettella[2], Christine L Brandt[2], Toni-Kim Clarke[26], David Coynel[31,34], Franziska Degenhardt[28,29], Srdjan Djurovic[2,37], Sarah Eisenacher[1], Matthias Fastenrath[31,34], Helena Fatouros-Bergman[11], Andreas J Forstner[28,29,38,39,40], Josef Frank[35], Francesco Gambi[41], Barbara Gelao[3], Leo Geschwind[30,31], Massimo di Giannantonio[41,42], Annabella Di Giorgio[3,43], Catharina A Hartman[44], Stefanie Heilmann-Heimbach[28,29], Stefan Herms[28,29,45], Pieter J Hoekstra[46], Per Hoffmann[28,29,45], Martine Hoogman[5,18], Erik G Jönsson[4,11], Eva Loos[31,34], Eleonora Maggioni[3,17], Jaap Oosterlaan[47], Marco Papalino[3], Antonio Rampino[3], Liana Romaniuk[26], Pierluigi Selvaggi[3,48], Gianna Sepede[3,41], Ida E Sønderby[2], Klara Spalek[31,34], Jessika E Sussmann[26], Paul M Thompson[49], Alejandro Arias Vasquez[21], Christian Vogler[30,31], Heather Whalley[26] [37]Department of Medical Genetics, Oslo University Hospital, Oslo, Norway. [38]Human Genomics Research Group, Department of Biomedicine, University of Basel, Basel, Switzerland. [39]Department of Psychiatry (UPK), University of Basel, Basel, Switzerland. [40]Institute of Medical Genetics and Pathology, University Hospital Basel, Basel, Switzerland. [41]Department of Neuroscience, Imaging and Clinical Sciences "G. D'Annunzio" University Chieti-Pescara, Pescara, Italy. [42]Department of Mental Health, National Health Trust, Chieti, Italy. [43]Fondazione Casa Sollievo della Sofferenza IRCCS San Giovanni Rotondo (FG), San Giovanni Rotondo, Italy. [44]Department of Psychiatry, Interdisciplinary Center Psychopathology and Emotion regulation, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [45]Department of Biomedicine & Institute of Medical Genetics and Pathology, Human Genomics Research Group and Division of Medical Genetics, Department of Biomedicine, University and University Hospital Basel, Basel, Switzerland. [46]Department of Child and Adolescent Psychiatry, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [47]Emma Children's Hospital, Academic Medical Center, Amsterdam, The Netherlands. [48]Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [49]Imaging Genetics Center, Stevens Institute for Neuroimaging & Informatics, University of Southern California, Los Angeles, CA, USA.

Karolinska Schizophrenia Project (KaSP) Consortium

Farde L[11], Flyckt L[11], Engberg G[50], Erhardt S[50], Fatouros-Bergman H[11], Cervenka S[11], Schwieler L[50], Agartz I[2,11,12], Collste K[11], Victorsson P[11], Malmqvist A[50], Hedberg M[50], Orhan F[50] [50]Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden

Conflicts of interest

A.M.-L. has received consultant fees from Blueprint Partnership, Boehringer Ingelheim, Daimler und Benz Stiftung, Elsevier, F. Hoffmann-La Roche, ICARE Schizophrenia, K. G. Jebsen Foundation, L.E.K Consulting, Lundbeck International Foundation (LINF), R. Adamczak, Roche Pharma, Science Foundation, Synapsis Foundation – Alzheimer Research Switzerland, System Analytics, and has received lectures including travel fees from Boehringer Ingelheim, Fama Public Relations, Institut d'investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Janssen-Cilag, Klinikum Christophsbad, Göppingen, Lilly Deutschland, Luzerner Psychiatrie, LVR Klinikum Düsseldorf, LWL PsychiatrieVerbund Westfalen-Lippe, Otsuka Pharmaceuticals, Reunions i Ciencia S. L., Spanish Society of Psychiatry, Südwestrundfunk Fernsehen, Stern TV, and Vitos Klinikum Kurhessen. J.K.B. has been in the past 3 years a consultant to / member of advisory board of / and/or speaker for Roche, Medice and Servier. He is not an employee of any of these companies, and not a stock shareholder of any of these companies. He has no other financial or material support, including expert testimony, patents, royalties. A.B. is a stockholder of Roche and has received lecture fees from Otsuka. M.Z. has

received unrestricted scientific grants from German Research Foundation (DFG), and Servier; further speaker and travel grants were provided by Otsuka, Servier, Lundbeck, Roche, Ferrer and Trommsdorff. S.C. has received grant support from AstraZeneca as a co-investigator, and has served as a one-off speaker for Otsuka-Lundbeck and Roche Pharmaceuticals. SCs spouse is an employee of SOBI pharmaceuticals. G.P. was an academic supervisor of a Hoffmann-La Roche collaboration grant (years 2015-16). B.F. has received educational speaking fees from Shire and Medice. All other authors declare no potential conflicts of interest.

Supplementary Information accompanies this paper at (https://doi.org/10.1038/s41398-018-0225-4).

Received: 4 July 2018 Accepted: 16 July 2018
Published online: 17 January 2019

References

1. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.* **30**, 67–76 (2008).
2. Ross, C. A., Margolis, R. L., Reading, S. A., Pletnikov, M. & Coyle, J. T. Neurobiology of schizophrenia. *Neuron* **52**, 139–153 (2006).
3. Lewis, D. A. & Lieberman, J. A. Catching up on schizophrenia: natural history and neurobiology. *Neuron* **28**, 325–334 (2000).
4. Shepherd, A. M., Laurens, K. R., Matheson, S. L., Carr, V. J. & Green, M. J. Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. *Neurosci. Biobehav. Rev.* **36**, 1342–1356 (2012).
5. Okada, N. et al. Abnormal asymmetries in subcortical brain volume in schizophrenia. *Mol. Psychiatry* **21**, 1460–1466 (2016).
6. Gupta, C. N. et al. Patterns of Gray Matter Abnormalities in Schizophrenia Based on an International Mega-analysis. *Schizophr. Bull.* **41**, 1133–1142 (2015).
7. van Erp, T. G. et al. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol. Psychiatry* **21**, 585 (2016).
8. Honea, R., Crow, T. J., Passingham, D. & Mackay, C. E. Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. *Am. J. Psychiatry* **162**, 2233–2245 (2005).
9. Haijma, S. V. et al. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophr. Bull.* **39**, 1129–1138 (2013).
10. Glahn, D. C. et al. Meta-analysis of gray matter anomalies in schizophrenia: application of anatomic likelihood estimation and network analysis. *Biol. Psychiatry* **64**, 774–781 (2008).
11. Ellison-Wright, I., Glahn, D. C., Laird, A. R., Thelen, S. M. & Bullmore, E. The anatomy of first-episode and chronic schizophrenia: an anatomical likelihood estimation meta-analysis. *Am. J. Psychiatry* **165**, 1015–1023 (2008).
12. Cooper, D., Barker, V., Radua, J., Fusar-Poli, P. & Lawrie, S. M. Multimodal voxel-based meta-analysis of structural and functional magnetic resonance imaging studies in those at elevated genetic risk of developing schizophrenia. *Psychiatry Res.* **221**, 69–77 (2014).
13. Bora, E. et al. Neuroanatomical abnormalities in schizophrenia: a multimodal voxelwise meta-analysis and meta-regression analysis. *Schizophr. Res.* **127**, 46–57 (2011).
14. Moberget, T, et al. Cerebellar volume and cerebellocerebral structural covariance in schizophrenia: a multisitemega-analysis of 983 patients and 1349 healthy controls. *Mol Psychiatry.* **23**, 1512–1520 (2018).
15. Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B. & Marquand, A. F. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci. Biobehav. Rev.* **57**, 328–349 (2015).
16. Doan, N. T. et al. Distinct multivariate brain morphological patterns and their added predictive value with cognitive and polygenic risk scores in mental disorders. *Neuroimage Clin.* **15**, 719–731 (2017).
17. Skatun, K. C. et al. Consistent Functional Connectivity Alterations in Schizophrenia Spectrum Disorder: A Multisite Study. *Schizophr. Bull.* **43**, 914–924 (2017).
18. Plis, S. M. et al. Deep learning for neuroimaging: a validation study. *Front. Neurosci.* **8**, 229 (2014).
19. Sabuncu, M. R., Konukoglu, E. & Alzheimer's Disease Neuroimaging, I. Clinical prediction from structural brain MRI scans: a large-scale empirical study. *Neuroinformatics* **13**, 31–46 (2015).
20. Rozycki, M. et al. Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. *Schizophr. Bull.* **44**, 1035–1044 (2018).
21. Chekroud, A. M. Bigger Data, Harder Questions-Opportunities Throughout Mental Health Care. *JAMA Psychiatry* **74**, 1183–1184 (2017).
22. Koutsouleris, N. et al. Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. *Brain* **138**, 2059–2073 (2015).
23. Schnack, H. G. 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* **84**, 299–306 (2014).
24. Salvador, R. et al. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLoS One* **12**, e0175683 (2017).
25. Owens, D. G. & Johnstone, E. C. Precursors and prodromata of schizophrenia: findings from the Edinburgh High Risk Study and their literature context. *Psychol. Med.* **36**, 1501–1514 (2006).
26. West, S. A. et al. The comorbidity of attention-deficit hyperactivity disorder in adolescent mania: potential diagnostic and treatment implications. *Psychopharmacol. Bull.* **31**, 347–351 (1995).
27. Wingo, A. P. & Ghaemi, S. N. A systematic review of rates and diagnostic validity of comorbid adult attention-deficit/hyperactivity disorder and bipolar disorder. *J. Clin. Psychiatry* **68**, 1776–1784 (2007).
28. Klassen, L. J., Katzman, M. A. & Chokka, P. Adult ADHD and its comorbidities, with a focus on bipolar disorder. *J. Affect Disord.* **124**, 1–8 (2010).
29. Chang, K. D. Course and impact of bipolar disorder in young patients. *J. Clin. Psychiatry* **71**, e05 (2010).
30. Consortium C-DGotPG.. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
31. Forstner, A. J. et al. Identification of shared risk loci and pathways for bipolar disorder and schizophrenia. *PLoS ONE* **12**, e0171595 (2017).
32. Owen, M. J. Intellectual disability and major psychiatric disorders: a continuum of neurodevelopmental causality. *Br. J. Psychiatry* **200**, 268–269 (2012).
33. Frangou, S., Schwarz, E. & Meyer-Lindenberg, A. Imagemend. Identifying multimodal signatures associated with symptom clusters: the example of the IMAGEMEND project. *World Psychiatry* **15**, 179–180 (2016).
34. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179–194 (1999).
35. Ashburner, J. & Friston, K. J. Voxel-based morphometry--the methods. *Neuroimage* **11**, 805–821 (2000).
36. Fernandez-Delgado, M., Cernadas, E., Barro, S. & Amorim,D. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15**, 3133–3181 (2014).
37. Fischl, B. Automatically Parcellating the Human Cerebral Cortex. *Cereb. Cortex* **14**, 11–22 (2004).
38. Desikan, R. S. et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).
39. Fischl, B. et al. Whole Brain Segmentation. *Neuron* **33**, 341–355 (2002).
40. Ashburner, J. A fast diffeomorphic image registration algorithm. *Neuroimage* **38**, 95–113 (2007).
41. Tzourio-Mazoyer, N. et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273–289 (2002).
42. Ho, D. E., Imai, K., King, G. Stuart, E. A.. MatchIt: nonparametric preprocessing for parametric causal inference. *J. Statist. Softw.* **42**, 1–28 (2011).
43. Yoav Benjamini, Y. H. Controlling the False Discovery Rate: a Practical and Powerful. Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).s
44. Robin, X. et al. pROC: an open-source package for R and S+to analyze and compare ROC curves. *BMC Bioinforma.* **12**, 77 (2011).
45. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

46. Wiener, A. La. M. Classification and Regression by randomForest. *R. News* **2**, 18–22 (2002).
47. Diaz-Uriarte R dAS. Variable selection from random forests: application to gene expression data. *Arxiv preprint q-bio/0503025* 2005.
48. Menze, B. H. et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinforma.* **10**, 213 (2009).
49. Diaz-Uriarte, R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinforma.* **8**, 328 (2007).
50. Corinna Cortes, V. V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
51. David Meyer E. D., KurtHornik, Andreas Weingessel and Friedrich Leisch. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071. 2017.
52. International Schizophrenia, C. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
53. Whitford, T. J. et al. Grey matter deficits and symptom profile in first episode schizophrenia. *Psychiatry Res.* **139**, 229–238 (2005).
54. Whitford, T. J. et al. Progressive grey matter atrophy over the first 2-3 years of illness in first-episode schizophrenia: a tensor-based morphometry study. *Neuroimage* **32**, 511–519 (2006).
55. Lieberman, J. A. et al. Antipsychotic drug effects on brain morphology in first-episode psychosis. *Arch. Gen. Psychiatry* **62**, 361–370 (2005).
56. Lo, M. T. et al. Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nat. Genet.* **49**, 152–156 (2017).
57. V Anttila B. B.-S., et al. Analysis of shared heritability in common disorders of the brain. *bioRxivorg* 101101/048991 2016.
58. Vernon, A. C. et al. Contrasting effects of haloperidol and lithium on rodent brain structure: a magnetic resonance imaging study with postmortem confirmation. *Biol. Psychiatry* **71**, 855–863 (2012).
59. Vernon, A. C., Natesan, S., Modo, M. & Kapur, S. Effect of chronic antipsychotic treatment on brain structure: a serial magnetic resonance imaging study with ex vivo and postmortem confirmation. *Biol. Psychiatry* **69**, 936–944 (2011).