Multisite Machine Learning Analysis Provides a Robust Structural Imaging Signature of Schizophrenia Detectable Across Diverse Patient Populations and Within Individuals

Martin Rozycki1,8, Theodore D. Satterthwaite1,2,8, Nikolaos Koutsouris3, Guray Erus1, Jimit Doshi1, Daniel H. Wolf1,2, Yong Fan1, Raquel E. Gur2, Ruben C. Gur2, Eva M. Meisenzahl4, Chuanjun Zhuo5, Hong Ying6, Hao Yan7, Weihua Yue7, Dai Zhang7, and Christos Davatzikos*,1

1Center for Biomedical Image Computing and Analytics, Department of Radiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 2Brain Behavior Laboratory, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 3Department of Psychiatry and Psychotherapy, Ludwig-Maximilians-University, Munich, Germany; 4Department of Psychiatry and Psychotherapy, Heinrich Heine University Dusseldorf, Dusseldorf, Germany; 5Tianjin Anning Hospital, Tianjin, China; 6Department of Radiology, Xijing Hospital, Fourth Military Medical University, Xi’an, China; 7Institute of Mental Health, Key Laboratory of Mental Health, Ministry of Health & National Clinical Research Center for Mental Disorders, Sixth Hospital, Peking University, Beijing, China

8These authors contributed equally to the article.

*To whom correspondence should be addressed; University of Pennsylvania, Richards Building, 7th Floor, 3700 Hamilton Walk, Philadelphia, PA 19104; tel: 215-746-4067, fax: 215-746-4060, e-mail: christos@rad.upenn.edu

Past work on relatively small, single-site studies using regional volumetry, and more recently machine learning methods, has shown that widespread structural brain abnormalities are prominent in schizophrenia. However, to be clinically useful, structural imaging biomarkers must integrate high-dimensional data and provide reproducible results across clinical populations and on an individual person basis. Using advanced multi-variate analysis tools and pooled data from case–control imaging studies conducted at 5 sites (941 adult participants, including 440 patients with schizophrenia), a neuroanatomical signature of patients with schizophrenia was found, and its robustness and reproducibility across sites, populations, and scanners, was established for single-patient classification. Analyses were conducted at multiple scales, including regional volumes, voxelwise measures, and complex distributed patterns. Single-subject classification was tested for single-site, pooled-site, and leave-site-out generalizability. Regional and voxelwise analyses revealed a pattern of widespread reduced regional gray matter volume, particularly in the medial prefrontal, temporolimbic and peri-Sylvian cortex, along with ventricular and pallidum enlargement. Multivariate classification using pooled data achieved a cross-validated prediction accuracy of 76% (AUC = 0.84). Critically, the leave-site-out validation of the detected schizophrenia signature showed accuracy/AUC range of 72–77%/0.73–0.91, suggesting a robust generalizability across sites and patient cohorts. Finally, individualized patient classifications displayed significant correlations with clinical measures of negative, but not positive, symptoms. Taken together, these results emphasize the potential for structural neuroimaging data to provide a robust and reproducible imaging signature of schizophrenia. A web-accessible portal is offered to allow the community to obtain individualized classifications of magnetic resonance imaging scans using the methods described herein.

Key words: psychosis/neuroimaging/MRI/brain/biomarker

Introduction

Schizophrenia is an often devastating illness that results in substantial morbidity and mortality worldwide.1 Structural brain abnormalities were noted by early anatomists,2,3 and results from in vivo magnetic resonance imaging (MRI) have supported the notion that schizophrenia is a brain disorder.4,5 Despite such advances, neuroimaging is typically not used as part of standard clinical care of psychotic disorders, in part due to heterogeneous results. Advances in clinical practice require robust imaging biomarkers of disease that can be used for diagnosis, determination of prognosis, and integration within trials of novel therapeutics.
As imaging technology has advanced and costs have declined, both the sample sizes and number of studies examining structural brain abnormalities in schizophrenia have expanded dramatically. Results from small single-site studies have often been heterogeneous, leading investigators to integrate findings across studies using meta-analyses of published data from hundreds of studies and thousands of patients. These studies provide increasingly consistent evidence of widespread patterns of structural deficits in schizophrenia, including gray matter loss that is prominent in the insular, cingulate, prefrontal, and temporal cortex.

However, such retrospective, literature-based meta-analyses are inherently limited by methodological variation. Scanner effects, processing choices, and analytic strategy can have a dramatic impact on results of individual studies, adding noise to results. The ENIGMA consortium was created to overcome such obstacles through the use of prospective meta-analyses, where all sites perform the same data processing, quality assurance, and group-level analyses. Using this approach, the ENIGMA-schizophrenia working group recently examined subcortical volumetric abnormalities in a sample of over 2000 patients and a similar number of controls, and reported volume reductions of small-moderate effect size (Cohen’s d: 0.2–0.4). These findings have subsequently been replicated by the COCORO consortium, which used a similar approach. Despite the enormous sample size provided by ENIGMA, such regional analyses ultimately provide a coarse-grained account of structural brain abnormalities associated with schizophrenia, and at present are limited to the subcortex. In contrast, one large recent study directly pooled data in a voxel-based, whole-brain mega-analysis, and delineated widespread loss of gray matter density that was maximal in insular cortex.

Importantly, despite accelerating efforts to pool structural imaging data to study psychosis, such meta-and mega-analyses have focused on mass-univariate approaches applied at the group level. Alternatively, brain-wide multivariate neuroanatomical patterns can be used to classify individuals as cases or controls using machine learning methods. By concisely summarizing high-dimensional data, multivariate classification has clear advantages for clinical translation. Single-site studies over the past decade have demonstrated that multivariate analyses can accurately discriminate patients and controls, as well as predict progression. One recent meta-analysis of single-site multivariate classification studies reported pooled classification sensitivity of 76%, with 79% specificity. However, no study to date has demonstrated single-subject classification in a large, multi-site, and multi-ethnic setting. In particular, it remains unknown whether classifiers trained on data acquired at one site will perform similarly on data acquired from a new site that was not included in the training set. Any practical clinical application would require an imaging biomarker that is robust to site differences, so that algorithms developed at academic centers could be applied in the community.

Here, we report results from a large-scale analysis of structural MRI data from 941 participants across 5 sites. Importantly, by pooling images instead of derived measures, and uniformly processing all data via identical pipelines, we were able to harmonize structural measurements and attenuate intersite differences. Building on prior efforts, our first hypothesis was that widespread structural brain abnormalities in adult patients with schizophrenia would be detectable across all cohorts. Our second hypothesis was that a reproducible neuroanatomical signature of schizophrenia would be present across sites and diverse populations, and be detectable at the individual patient level. Our approach included mass-univariate analyses of regional volumes defined using state-of-the-art multi-atlas segmentation, optimally discriminative voxel based analyses (ODVBA) for local-multivariate pattern analysis, and individual level classification using consensus-based machine learning.

Methods

Participants and Image Acquisition

Data from 5 MRI studies were included in the current multi-site mega-analysis. All subjects were part of previous studies overseen by local institutional policies. Sample demographics of these data sets are given in table 1. Detailed MRI acquisition protocol information for each site is given in supplementary table 1. Anonymized data provided by the 5 sites were provided to and analyzed by the University of Pennsylvania’s Center for Biomedical Image Computing and Analytics. Raw MR scans were initially examined for motion, image artifacts, or restricted field-of-view. Scans were also checked for lesions, but this population of patients being relatively young, there were few subjects with lesions, and the total lesion load of these subjects was low. Further details of quality control procedures are given in supplementary methods.

Image Processing

Following magnetic field inhomogeneity correction, MR images were first segmented into a hierarchically organized set of 259 anatomical regions, ranging from total brain volume down to individual cortical gyri and deep structures. To parcellate the brain into anatomically defined regions, we used a highly accurate multi-atlas consensus labeling procedure, which was recently the top-ranked method at an independent international competition. Besides regional volumes, we also calculated a set of rich regional descriptors based on intensity, shape, size, and texture properties of each region, to be used as an additional set of input features in subsequent multivariate classification. See supplementary methods for more details.
Table 1. Sample Demographics

<table>
<thead>
<tr>
<th>Group, No. (%)</th>
<th>Harmonized (n = 835)</th>
<th>SET1 (n = 325)</th>
<th>SET2 (n = 180)</th>
<th>SET3 (n = 330)</th>
<th>SET4 (n = 42)</th>
<th>SET5 (n = 64)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Controls</td>
<td>Patients</td>
<td>Controls</td>
<td>Patients</td>
<td>Controls</td>
<td>Patients</td>
</tr>
<tr>
<td></td>
<td>448 (53.7)</td>
<td>387 (46.3)</td>
<td>170 (52.3)</td>
<td>155 (47.7)</td>
<td>105 (58.3)</td>
<td>75 (41.7)</td>
</tr>
<tr>
<td>Female sex, no. (%)</td>
<td>215 (48.0)</td>
<td>148 (38.2)</td>
<td>100 (58.8)</td>
<td>70 (45.2)</td>
<td>61 (58.1)</td>
<td>39 (52.0)</td>
</tr>
<tr>
<td>Age, mean (SD)</td>
<td>33.9 (11.6)</td>
<td>35.3 (13.5)</td>
<td>35.7 (12.0)</td>
<td>39.0 (8.7)</td>
<td>35.0 (8.7)</td>
<td>35.5 (9.1)</td>
</tr>
<tr>
<td>ICV, mean (SD), cm³</td>
<td>1415.8 (148.5)</td>
<td>1398.2 (145.4)</td>
<td>1410.8 (166.5)</td>
<td>1378.8 (149.3)</td>
<td>1382.3 (122.9)</td>
<td>1361.6 (153.1)</td>
</tr>
<tr>
<td>CPZ dosage, mean (min–max)</td>
<td>N/A</td>
<td>436.0 (0.0–2766.7)</td>
<td>N/A</td>
<td>506.8 (0.0–2766.7)</td>
<td>N/A</td>
<td>478.0 (0.0–1775.0)</td>
</tr>
<tr>
<td>Duration of illness, mean (min–max), years</td>
<td>N/A</td>
<td>10.78 (0.00–43.90)</td>
<td>N/A</td>
<td>18.21 (0.60–43.90)</td>
<td>N/A</td>
<td>9.58 (0.00–27.00)</td>
</tr>
<tr>
<td>Clinical symptom scores, no. (%), type</td>
<td>N/A</td>
<td>357 (92.2), Various</td>
<td>N/A</td>
<td>143 (92.3), SAPS/SANS</td>
<td>N/A</td>
<td>59 (78.7), PANSS</td>
</tr>
<tr>
<td>Acquisition protocol</td>
<td>Various</td>
<td>MPRAGE</td>
<td>BRAVO</td>
<td>MPRAGE</td>
<td>MPRAGE</td>
<td>MPRAGE</td>
</tr>
<tr>
<td>Field strength</td>
<td>Various</td>
<td>University of Pennsylvania, Philadelphia, PA</td>
<td>3 T University of Pennsylvania, Philadelphia, PA</td>
<td>Tianjin Anning Hospital, Tianjin, China</td>
<td>1.5 T Ludwig-Maximilian-University, Munich, Germany</td>
<td>3 T Xijing Hospital, Xi'an, China</td>
</tr>
<tr>
<td>Location</td>
<td>Various</td>
<td>3 T University of Pennsylvania, Philadelphia, PA</td>
<td>3 T Tianjin Anning Hospital, Tianjin, China</td>
<td>1.5 T Ludwig-Maximilian-University, Munich, Germany</td>
<td>3 T Xijing Hospital, Xi'an, China</td>
<td>3 T Peking University, Beijing, China</td>
</tr>
</tbody>
</table>

Note: N/A = not applicable.

*CPZ dosage equivalents available for 71 of SET1 patients, 59 of SET2 patients, and 125 of SET3 patients.

*Duration of illness available for 130 of SET1 patients, 59 of SET2 patients, and 153 of SET3 patients.
To perform high-resolution voxel-based analysis, each image’s gray matter segmentation was obtained using a previously reported and validated method, MICO. Segmented maps were warped into a common template space using a highly accurate deformable registration (DRAMMS), yielding regional volumetric maps (RAVENS) that quantify gray matter volume at each voxel. We masked out subcortical gray matter structures on RAVENS maps, because available structural T1-weighted images do not provide sufficient contrast for reliable intensity-based segmentation; analysis of subcortical structures was limited to the more robust atlas-based regional parcellation described above.

Intersite Image Harmonization
Scanner and demographic differences across MRI samples offer challenges for combining data. Accordingly, we harmonized data across sites by estimating intracranial volume (ICV), site, age, and sex effects on each imaging measure within a pooled sample of controls using a linear model. The coefficients estimated from this model were then applied to the whole sample including patient data. Effectively, this removed the influence of site and demographic effects on the difference between patients and controls. Importantly, this control-based harmonization model was always cross-validated, if it was only derived from the training set, and subsequently applied to the test set. Due to the small sample-sizes of the SET4 and SET5 sites, these sites were not included in the harmonization procedure and were set aside for use as validation cohorts in the classification experiments described later. The harmonization procedure was applied independently on each input image feature, including regional features and voxelwise RAVENS maps.

Group-Level Statistical Analyses
All group-level analyses were conducted within the harmonized cohort (n = 835). Regional case-control comparisons were conducted using 2 sample t-tests. Results are reported as effect sizes (Cohen’s d-statistic). Effect of medication and disease duration were each evaluated separately using linear regression. Voxelwise analyses of RAVENS maps used ODVBA, a local-multivariate pattern analysis method that has been shown in prior studies to have greater sensitivity than both standard voxel-based morphometry and searchlight-based local-multivariate analyses. Type I error was controlled using the false discovery rate (FDR) correction (Q < 0.05). Voxelwise effect sizes were reported in regions that were found to be significant by ODVBA.

Multivariate Classification
The above analyses provide a description of structural brain differences between schizophrenia and control groups. As a final step, we used machine learning techniques to create a multivariate classifier that provides individual-level prediction of group status. We used a linear support vector machine (SVM) classifier, a widely used method with extensive validation in various similar high dimensional pattern classification problems. The input features included voxelwise values from RAVENS maps, regional volumes, and an extended set of shape, intensity and texture features computed for each region. A classifier was run separately for each feature type using 10-fold cross-validation. A consensus prediction was then obtained by combining the predictions of each classifier through a weighted averaging based on their cross-validated individual performances. This procedure was performed using an external 10-fold cross-validation (supplementary figure 4), except in experiments that use independent training and testing sets.

To maximize generalizability, we applied the SVM classification using default parameters (C = 1), without feature selection. Optimization of such parameters would likely further improve accuracy, albeit at the risk of overfitting the data.) See supplementary methods for further details.

Multivariate models were trained under 3 different conditions to evaluate the impact of pooling data across sites and training/testing on different sites. First, models were trained and tested separately within each site using 10-fold cross-validation. Second, a single 10-fold cross-validated SVM classification analysis was conducted using the complete sample of harmonized pooled data. Third, we performed training/testing on different sites through leave-one-site-out cross-validation. This procedure included applying a model trained on the complete pooled data to the 2 independent validation cohorts. This procedure aimed to evaluate the generalizability of the classifier to a totally new data set sampled from an entirely different population and scanner. Importantly, in this procedure harmonization was done only within the training set, and the control-based regression model was applied to the test set. Before training the classifier, the training and testing sets were z-scored separately, with the assumption that the test set has a similar patient to control ratio as the training set. For all analyses, individual-subject classification performance was evaluated using receiver operating characteristic (ROC) curves. Classification performance was summarized using both the area under the curve (AUC) and the classification accuracy metrics.

Finally, to assess the clinical relevance of the MRI-based classification, we investigated correlations between the classifier output and clinical symptom scores of individual subjects. The symptom scores for different sites were obtained using common symptom rating scales in schizophrenia research: the Scale for the Assessment of Positive Symptoms, the Scale for the Assessment of Negative Symptoms, and the Positive And Negative Syndrome Scale. Table 1 details the available type of scores for each site. The scores were normalized by z-scoring them within each site and pooled together. For this analysis, individualized pseudo-probabilities of having the neuroanatomical
signature of schizophrenia were estimated from the classifier’s output using sigmoid fits. These pseudo-probabilities correspond to the degree to which a given individual’s brain appears to match the schizophrenia pattern.

Results

Regional and Voxelwise Analyses Reveal Distributed Structural Brain Abnormalities in Schizophrenia

Mass-univariate regional analyses of pooled data delineated robust abnormalities of brain structure that survived FDR correction. The structures that show the most significant group differences, both in positive and negative directions, are shown in figure 1. Patients with schizophrenia had marked ventricular expansion, as well as larger pallidum volumes. Cortical gray matter volume loss was also quite evident, with maximal effects in the prefrontal cortex (superior frontal gyrus; \( d = 0.51 \)) as well as temporal cortex, parietal cortex, insula, and amygdala. The effect sizes of volumetric changes were strongly correlated across sites (pairwise correlation coefficient range: 0.71–0.81), indicating strong consistency despite protocol, scanner, and demographic differences across sites. A complete list of all regions with significant group differences (\( q < 0.05 \)) is provided in supplementary table 1. Site-specific analysis results are provided in supplementary table 3.

High-resolution voxelwise analyses of harmonized RAVENS maps in the pooled sample using ODVBA delineated a distributed pattern of volume loss (figure 2), and showed good agreement with the results of the ROI-based analysis (supplementary figure 1). The results of the ODVBA analyses performed using each data set separately are shown in supplementary figure 2. The effects found for individual data sets were consistent with each other and with the pooled analysis, but less significant than the pooled analysis results, emphasizing the enhanced power of the pooled analysis compared to the smaller single-site analyses.

We also evaluated the impact of disease duration and antipsychotic load (chlorpromazine equivalents\(^5\)) on the regional data (supplementary table 4). In contrast to the very robust and distributed group differences, medication effects were only seen in the frontal operculum, where higher medication dose was associated with lower volume. Duration of illness was associated with larger lateral ventricles and pallidum, as well as reduced volume of regions including the middle frontal gyrus, parahippocampal gyrus, and hippocampus.

Multivariate Classification is Accurate Using Data Pooled Across Sites

Classification results are detailed in table 2 for the within-site, pooled, and between-site experiments. Compared to individual within-site classification, pooled classification resulted in higher classification accuracy for each of SET1, SET2, and SET3, with 0.02 to 0.045 increase in AUC. The ROC curve for the cross-validated classification within the pooled-data is shown in figure 3a. These results indicate that pooling data across sites is not only viable, but it actually improves the classification results. The leave-site-out cross-validation procedure on SET1, SET2 and SET3, ie, using models trained exclusively on data collected at different sites, achieved AUC values between 0.78 and 0.80, which were comparable to pooled or within site cross-validation results. Besides, combination of training data from 2 separate external sites consistently outperformed training on each external site individually, suggesting...
that pooling a larger number of data sets for training helps to increase accuracy and robustness of the classifier. Classification results for the independent validation cohorts, SET4 and SET5 were also high (accuracies 0.738 and 0.774), further emphasizing the feasibility of using models trained exclusively on data collected at a different site.

### Discussion

In a large-scale analysis of data pooled across sites, we demonstrated that the widespread structural brain abnormalities associated with schizophrenia can be used for accurate single-subject classification using machine learning techniques. Application of state-of-the-art consensus-based classifiers on different types of imaging features yielded high performance, and indicated that multivariate

### Classification Probability Is Related to Negative Symptoms

We found a significant positive correlation ($r = .228$, $P = .00003$) between the individual pseudoprobabilities of having neuroanatomical pattern and the negative symptom scores of the subjects (figure 3b). No significant correlations were found between the positive symptom scores and the classification pseudoprobabilities, nor was there a significant relationship to the duration of illness.

### Table 2. Classification Results

<table>
<thead>
<tr>
<th>Train Site</th>
<th>Test Site, AUC/Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET1</td>
<td>0.767/0.696</td>
</tr>
<tr>
<td>SET2</td>
<td>0.770/0.720</td>
</tr>
<tr>
<td>SET3</td>
<td>0.782/0.701</td>
</tr>
<tr>
<td>Other sites</td>
<td>0.785/0.723</td>
</tr>
<tr>
<td>Pooled</td>
<td>0.813/0.736</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Train Site</th>
<th>Test Site, AUC/Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET1</td>
<td>0.765/0.694</td>
</tr>
<tr>
<td>SET2</td>
<td>0.872/0.808</td>
</tr>
<tr>
<td>SET3</td>
<td>0.793/0.706</td>
</tr>
<tr>
<td>Other sites</td>
<td>0.805/0.739</td>
</tr>
<tr>
<td>Pooled</td>
<td>0.893/0.803</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Train Site</th>
<th>Test Site, AUC/Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET1</td>
<td>0.801/0.707</td>
</tr>
<tr>
<td>SET2</td>
<td>0.791/0.710</td>
</tr>
<tr>
<td>SET3</td>
<td>0.802/0.732</td>
</tr>
<tr>
<td>Other sites</td>
<td>0.804/0.731</td>
</tr>
<tr>
<td>Pooled</td>
<td>0.829/0.762</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Train Site</th>
<th>Test Site, AUC/Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET1</td>
<td>0.588/0.643</td>
</tr>
<tr>
<td>SET2</td>
<td>0.852/0.810</td>
</tr>
<tr>
<td>SET3</td>
<td>0.639/0.524</td>
</tr>
<tr>
<td>Other sites</td>
<td>0.639/0.524</td>
</tr>
<tr>
<td>Pooled</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Train Site</th>
<th>Test Site, AUC/Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET1</td>
<td>0.831/0.758</td>
</tr>
<tr>
<td>SET2</td>
<td>0.931/0.806</td>
</tr>
<tr>
<td>SET3</td>
<td>0.876/0.806</td>
</tr>
<tr>
<td>Other sites</td>
<td>0.912/0.774</td>
</tr>
<tr>
<td>Pooled</td>
<td>N/A</td>
</tr>
</tbody>
</table>

**Note:** Color codes show the set of results for the 3 different settings in classification experiments (blue: within site classification with leave 10% out cross-validation; green: classification with pooled data with leave 10% out cross-validation; red: classification with training/testing on different sites). N/A = not applicable.

*Sites 1, 2, and 3 comprised the discovery cohort. Sets 4 and 5 comprised the validation cohort.*

*For sites 1, 2, and 3, other sites indicates that 2 of the 3 sites were pooled for training, and the third site was used for testing. For sites 4 and 5, other sites indicates the pooled data from sites 1, 2, and 3 used for training.*

**Fig. 3.** Classification results. (a) ROC curve for patient vs control classification. The classification was performed using the pooled data and with leave-10%-out cross-validation through consensus-voting of all input feature sets, ie, RAVENS maps, regional volumes, and region-based descriptors. (b) Correlation between classifier output and negative clinical symptom scores (SANS for SET1, PANSS Negative for sets 2 and 3). Classification scores were obtained by converting the distances of the test samples from the discrimination hyperplane to pseudo-probabilities using sigmoid fits. Negative symptom scores were $z$-scored within data sets and then pooled together. The correlation for classifier score with negative symptom score was equal to $r = .228$ ($P = .00003$). No correlations were found between positive symptom scores and the classifier output ($r = -.019$, $P = .73$).
neuroanatomical signatures have the potential to become a quantitative imaging biomarker for schizophrenia. Importantly, these procedures remained accurate when trained and tested on data from completely different sites, populations, and scanners, highlighting the reproducibility of this imaging signature and, therefore, its translational potential.

**Structural Brain Abnormalities in Schizophrenia Are Widespread and Visible at Multiple Scales**

Capitalizing on a sample of nearly 1000 participants amassed via a multi-site, mega-analytic design, we provide robust evidence of widespread structural brain abnormalities in schizophrenia. These were present on multiple scales, and were evident in both analyses of regional volumes as well as high-resolution voxelwise analyses. Brain abnormalities in schizophrenia have been studied using brain imaging for 3 decades. While abnormalities have been consistently documented by single-site studies, the relative degree to which different brain regions were impacted has varied considerably by sample. Large-scale meta-analyses have pooled data, but they have been limited by important differences in data processing across studies. The ENIGMA and COROCO consortia, as well the study of Gupta et al., have overcome this limitation with rigorous meta-analyses of data processed in consistent fashion (albeit not always with identical pipelines harmonized on raw data), and recently provided evidence of subcortical volume loss of moderate effect size. Our analysis of subcortical structures provided convergent results to those provided by ENIGMA and other studies, including loss of hippocampus, thalamus, and amygdala volume; expansion of the pallidum; and ventricular enlargement. However, we build upon such findings from the subcortex and demonstrate cortical abnormalities of often large effect size, with a maximal deficit in the prefrontal cortex. In line with previous studies, duration of illness was associated with increasing expansion of the pallidum and ventricles. In contrast to the lack of medication effects found in ENIGMA, but in line with a previous large-scale retrospective meta-analysis, higher dose of antipsychotics was associated with larger pallidum and reduced gray matter volume in several frontal regions.

**Multivariate Classification of Individuals Is Accurate Even Using Data From Different Scanners and Sites**

While mass-univariate analyses provide a valuable description of the structural brain abnormalities associated with schizophrenia at the group level, they cannot function as biomarkers in individual patients. Accordingly, our primary focus was to use machine learning techniques to create a classifier that leveraged the complex multivariate pattern of structural deficits. The classifier we described here could differentiate patients and controls with a high degree of accuracy, with levels comparable to those previously reported in single-site studies.

A highly significant positive correlation was found between the classification output and negative symptom scores. Thus, the probability to which an individual’s multivariate pattern of structural anatomy was classified as consistent with schizophrenia was correlated with the burden of negative symptoms. In contrast, there was not a significant correlation with positive symptoms. This result is consistent with extensive prior research linking negative symptoms to structural brain abnormalities in schizophrenia.

Critical to this approach was that multivariate classifiers could retain accurate prediction even when trained and tested on data from separate scanners. Machine learning can be quite robust to intersite variations, by virtue of depending on the relative weights, ie, contrasts, of a complex pattern rather than on absolute volume-based thresholds, which can be impacted by slight variations of image contrast across scanners. These results suggest that models trained on data collected at specialized academic centers have the potential to be applied to data acquired in the community in order to calculate an imaging biomarker of schizophrenia. Importantly, leave-one-site-out was tested without the need for harmonization of the testing data with the training data set, but only via z-scoring within the new test site data. The classification pipeline is available in our image processing portal as a web-accessible application (CBICA Image Processing Portal: [https://ipp.cbica.upenn.edu/](https://ipp.cbica.upenn.edu/)). The web application allows users to submit single or multiple T1-weighted MRI images and outputs individualized scores, computed using automatically extracted imaging features and the presaved classification model from the harmonized training data. Thus, for any new data set, this web portal provides the scientific and clinical community with a freely accessible quantitative index of the imaging-based biomarker described here.
Limitations and Future Directions

Although this study benefited from a large sample size and advanced analytics, certain limitations should be noted. While developing a distinctive imaging signature of schizophrenia is an important endeavor, its value in predicting disease progression, treatment response, and in assisting differential diagnosis needs to be investigated in future studies. We note that the proposed imaging index, which can be calculated at an individual level using the provided online tool, cannot be used as a direct diagnostic index, as it has not been validated in heterogeneous clinical populations. Applications to samples that are not enriched for schizophrenia may result in a high false-positive rate. Like most other psychiatric conditions, schizophrenia is associated with substantial co-morbidity of other psychiatric conditions (such as mood disorders). Unfortunately, co-morbid conditions were not consistently recorded across the participating sites, and thus could not be evaluated in the present analysis. Future studies should evaluate co-morbidity of other psychiatric conditions on classification performance. Expanded data sets that seek to parse heterogeneity within diagnostic groups will be critical. 39,59–61

Recent work67 using heterogeneity analyses offers hope that semi-supervised machine learning methods can provide a better understanding of the heterogeneity of neuroanatomical signatures of schizophrenia. Finally, future work that combines large-scale imaging data sets in youth with well-characterized longitudinal clinical follow up with advanced multivariate analytics will be critical areas of focus moving forward.

Conclusions

Using a pooled mega-analytic strategy, the present data provide among the most robust evidence to date of structural brain abnormalities in adults with schizophrenia. Furthermore, these results emphasize that such signals can be used to derive highly accurate multivariate models that allow for discrimination at the level of individual patients, thereby providing a robust neuroanatomical signature of schizophrenia. Critically, this signature remains accurate even when the classifier is trained using data from different sites and scanners. Taken together, our findings highlight the accelerating promise of imaging-based biomarkers in major neuropsychiatric illnesses such as schizophrenia. The classifier used in this paper for classification is publicly available online at the CBICA Image Processing Portal.

Supplementary Material

Supplementary data are available at Schizophrenia Bulletin online.

Funding

The research at the University of Pennsylvania site was supported in part by the National Institute of Health (P50MH096891 to R.E.G., R01MH107703 to T.D.S., R01MH101111 to D.H.W) and the Dowshen Program for Neuroscience.

Acknowledgments

At the University of Pennsylvania site, thanks to Monica Calkins, PhD for assistance with clinical phenotyping.

Disclosures: R.E.G. participated in an advisory board for Otsuka Pharmaceuticals and RCG receives royalties from the Brain Resource Center and serves without compensation on an advisory board for Lumosity. All other authors have no disclosures.

References


50. Andreasen N. *The scale for the assessment of negative symptoms (SANS)*. Iowa City: University of Iowa; 1983.


