**Classification of first-episode psychosis using cortical thickness: a large multicenter MRI study.**

**Pigoni et al.**

**Supplementary Information**

**Supplementary Methods**

**1. Recruitment procedures and criteria**

Site 1: Santander

Data were acquired as part of a large prospective longitudinal study on first episode psychosis in the region of Cantabria(Tordesillas-Gutierrez et al., 2018). Individuals with a first episode psychosis (FEP) were recruited from both inpatient units and community services throughout the entire region. Patients were included if they met the following criteria: 1) age 15–60 years; 2) experiencing a FEP; 3) DSM-IV criteria for a diagnosis of schizophrenia, schizophreniform disorder, schizoaffective disorder, brief reactive psychosis, or not otherwise specified psychosis; and 4) no prior treatment with antipsychotic medication or, if previously treated, a total lifetime of adequate antipsychotic treatment of less than 6 weeks. Patients with DSM-IV based diagnoses of mental retardation or substance dependence (except nicotine dependence) were excluded. Healthy controls were recruited from the community through advertisements and were screened for current or past history of psychiatric, mental retardation, neurological or general medical illness, including substance dependence and significant loss of consciousness.

Site 2: Jena

Subjects were enrolled by trained and board-certified psychiatrists and diagnoses were made based on DSM-IV criteria and confirmed by standardized structured clinical interviews (SCID-IV). Exclusion criteria were (1) a history of previous psychotic disorder or manic episode, (2) substance induced psychotic disorder, (3) acute suicidal or aggressive behavior, (4) a current DSM-IV diagnosis of substance dependence, (5) neurological disorders (e.g., epilepsy), (6) IQ<70, (7) structural brain changes apparent on MRI scan, (8) previous treatment with an antipsychotic or mood stabilizing agent, and (9) any implanted metal or device that would be affected by the magnetic field of the scanner. Healthy controls were recruited from the community through advertisements and were screened for current or past history of psychiatric, mental retardation, neurological or general medical illness, including substance dependence and significant loss of consciousness.

Site 3: Prague

Patients were recruited at the Bohnice Psychiatric Hospital, Prague. The inclusion criteria were: (1) the diagnosis of schizophrenia or the diagnosis of an acute psychotic disorder, as made by a psychiatrist, according to the ICD-10 criteria; (2) the first episode of psychotic illness; (3) the duration of untreated psychosis = or < 24 months. The healthy control subjects were recruited via an advertisement from a similar sociodemographic background. The main exclusion criteria for the control subjects were a personal lifetime history of any psychiatric disorder, or any substance abuse, established by the Mini International Neuropsychiatric Interview (M.I.N.I.)(Sheehan et al., 1998) and a family history of a psychiatric illness in first- or second-degree relatives.

Site 4: London

Participants were recruited from the South London and Maudsley Foundation Trust and scanned at the Institute of Psychiatry, Psychology and Neuroscience in London (England). Diagnosis of schizophrenia was formulated by an experienced psychiatrist using the ICD-10 criteria. Healthy controls were recruited through local advertisement from the same geographical areas as patients. A screening tool (Psychosis Screening Questionnaire) was used to exclude the presence of psychotic symptomatology or a history of psychotic illness. Additional exclusion criteria for all participants included learning disabilities (based as an IQ < 70), current or past neurological illness, brain injury with loss of consciousness for more than 1 hour, and suspected or confirmed pregnancy.

Site 5: Basel

Participants were recruited in a specialized clinic for the early detection of psychosis at the University Hospital of Psychiatry, Basel, Switzerland. All patients were competent to give informed consent. They were able to understand relevant study information, including the reasons why they were being asked to participate and the procedures of the study, and they understood the consequences of accepting or declining the invitation to participate and how to discontinue their participation. Patients with FEP fulfilled the criteria for acute psychotic disorder according to the ICD-10 or DSM-IV. Exclusion criteria were the following: history of previous psychotic disorder, psychotic symptomatology secondary to an organic disorder, recent substance abuse according to ICD-10 research criteria, psychotic symptomatology associated with an affective psychosis or a borderline personality disorder, age younger than 18 years, inadequate knowledge of the German language, and IQ lower than 70. Healthy controls were recruited from the same geographical area as patients. Controls had to have no current psychiatric disorder; no history of psychiatric illness, head trauma, neurologic illness, serious medical or surgical illness or substance abuse; and no family history of any psychiatric disorder as assessed by an experienced psychiatrist in a detailed clinical assessment.

Site 6: Rome

All patients were diagnosed by senior clinical psychiatrists using the structured clinical interview for DSM-IV-TR (SCID-I/P)(Gorgens, 2011). Other inclusion criteria were: 1) age between 18 and 65 years; 2) onset of psychosis <24 months; 3) no dementia or cognitive deterioration according to the DSM-IV-TR criteria, and a Mini-Mental State Examination (MMSE)(Folstein et al., 1975) score higher than 24, consistent with normative data in the Italian population(Magni et al., 1996); and 4) suitability for a Magnetic Resonance Imaging (MRI) scan. Exclusion criteria were: 1) a history of alcohol or drug dependence or abuse in the last two years according to the DSM-IV-TR, 2) traumatic head injury, 3) any past or present major medical or neurological illness, 4) any other psychiatric disorder or mental retardation diagnosis and 5) MRI evidence of focal parenchymal abnormalities or cerebrovascular diseases. Healthy controls were screened for any current or past diagnosis of DSM-IV-TR axis I or II disorders using the SCID-I and SCID-II(Gorgens, 2011). A diagnosis of schizophrenia or any other mental disorder in first-degree relatives was also an exclusion criterion.

Site 7: Verona

The subjects were recruited in the frame of Psychosis Incident Cohort Outcome Study (PICOS), a multi-site naturalistic research in a large epidemiological area of the Veneto region, Italy(Lasalvia et al., 2012). Inclusion criteria, based on the over-inclusive screening methodology adopted in the WHO ten-country study(Jablensky et al., 1992), were: (1) age 15–54 years; (2) residence in the Veneto region; (3) presence of (a) at least one of the following symptoms: hallucinations, delusions, qualitative speech disorder, qualitative psychomotor disorder, and bizarre or grossly inappropriate behavior, or (b) at least two of the following symptoms: loss of interest, initiative and drive, social withdrawal, episodic severe excitement, purposeless destructiveness, overwhelming fear, and marked self-neglect; and (4) first lifetime contact with any mental health service located in PICOS area during the study period occasioned by symptoms enumerated in (3). The exclusion criteria were: (1) prior treatment with an antipsychotic agent for more than 3 months; (2) mental disorders due to a general medical condition; and (3) moderate to severe mental retardation. Additional exclusion criteria for MRI were adopted (i.e. pregnancy; metallic objects in the body). Healthy controls were recruited from the community through advertisements and were screened for current or past history of psychiatric, mental retardation, neurological or general medical illness, including substance dependence and significant loss of consciousness.

Site 8: Munich

Subjects with first acute psychotic episode according to DSMIV/ICD-10 criteria were included in the first week after admission to the hospital. History of the first episode of psychotic symptoms was confirmed by the psychiatrist admitting, and by close relatives. The inclusion criteria were (1) diagnosis of definite or probable schizophrenia; (2) admission for first episode of illness; (3) fewer than 24 weeks (lifetime) of neuroleptic treatment; (4) age between 18 and 50 years. Exclusion criteria were: included learning disabilities (based as an IQ < 70), current or past neurological illness, brain injury with loss of consciousness for more than 1 hour, and suspected or confirmed pregnancy. Healthy controls were recruited from the community through advertisements and were screened for current or past history of psychiatric, mental retardation, neurological or general medical illness, including substance dependence and significant loss of consciousness. Family history for psychiatric conditions was an additional exclusion criterion.

**2. Visual inspection of MRI reconstruction: criteria of exclusion**

A trained psychiatrist visually inspected each reconstructed image from anterior to posterior in coronal plane and from inferior to superior in axial plane, checking for bad reconstructions.

Similar to prior work(Rosen et al., 2018; Savalia et al., 2017), all images were rated on quality in the 3-class framework, ranging from images that suffer from gross artifacts and were considered unusable and labelled as “poor reconstruction”, to images free from visible artifact. The intermediate category was used for images with some artifact, but which would still be considered usable.

We took into account artifacts (i.e. movement artifacts), gyri that were not correctly included into the reconstruction as well as dural inclusions. As a rule, if more than 4 consecutive slices presented one artifact or a dural inclusion or a skipped gyrus, the reconstruction was labelled as “poor reconstruction”. From one to three consecutive slices presenting artifacts fell in the intermediate category.

**3. Machine learning models**

SVMs are multivariate statistical methods that provide optimal decision rules for classifying individuals. The algorithm maps the input data into a feature space using a set of similarity functions known as kernels. In this feature space, the model finds the optimal separating hyperplane by finding the largest margin of separation between the two classes within the training set. Once the hyperplane is determined, it can be used to predict the class of new unseen observations( Dwyer et al., 2018). In the recent years, they have been applied in a range of biomedical questions and for diagnostic purposes( Dwyer et al., 2018).

To allow for unbiased estimation of generalizability and prevent information leakage between the training and the validation models(Ruschhaupt et al., 2004), a double cycle, nested cross-validation (CV) framework(Koutsouleris et al., 2018, 2016) was built. In a k-fold CV, an entire test fold is then left out, while the model is learned on the rest of the training data, and then tested on the left-out individuals. This procedure is repeated for a prespecified number of k folds (in our analyses k=10) and results in stable estimates of generalizability because the training groups are more variable and there are more individuals in the left-out test sets. Moreover, nested CV includes a CV cycle within another, superordinate CV cycle that is ultimately used to assess the generalizability of the models ( Dwyer et al., 2018). In both inner and outer CV, a 10-fold CV cycle was employed. So, based on the nested scheme, in the outer-CV our sample was randomly divided into 10 folds and a fold was held-out as test set. The rest of the data went into the 'nest' where it was again subdivided into 10 folds, a test fold is held-out, and the rest of the data was used for training. This procedure was repeated so that every fold is held-out as a test set. Parsimonious models that contributed most to the discriminative pattern separating FEP and HC are created at the inner CV level. To enforce an unbiased estimation of classification generalizability, these models were then applied to the test data at the outer CV level, which consisted of study participants never used for training the discriminative algorithm. In other words, all model training steps that use group-level statistical procedures (e.g regress out of covariates effects, feature selection) occur only in the inner CV training data. The inner CV test data are used to choose the best hyperparameter combinations. Finally, the outer CV test data measure exclusively the models’ generalizability to unseen data. We extended nested CV to repeated nested CV at the outer cross-validation cycle by randomly permuting the participants within their groups (number of permutations = 10) and repeating the CV cycle for each of these permutations

Cross-Validation scheme:

 CV1 (inner cycle): 10 folds x 10 permutations

 CV2 (outer cycle): 10 folds x 10 permutation

Feature preprocessing:

Step 1: Pruning of non-informative columns from matrix [Zero Var, Nan, Inf]

Step 2: Partial correlations (Covariates: Gender, Age, Site, Mean Euler) were calculated on healthy controls (HC) and regressed out on the whole sample.

Given that HC and FEP significantly differed for gender (Supplementary Table 1), and given the effect of age on brain structures, we regressed out the variance associated with these variables. Specifically, we removed the variance associated with age and gender from the feature values within each inner and outer CV fold through partial correlations. The effect of nuisance covariates was calculated only on HC and then applied to the whole sample in each CV fold. This procedure allowed us to remove the “standard” effect of age and sex associated with the general population, while not removing any relevant information from the FEP group in relation to the disease. The same procedure was applied to sites and Euler number, in order to remove the heterogeneity associated with the different MRI scans employed in different centers(Schnack and Kahn, 2016) and the variance associated with the quality of the reconstruction(Rosen et al., 2018).

Step 3: PCA

The dimensionality of the training cases’ thickness values was reduced by means of Principal Component Analysis (PCA)(Hansen et al., 1999), a well-established unsupervised method for feature reduction in neuroimaging. PCA geometrically projects the data into lower dimensions called principal components (PCs) and applies an orthogonal transformation to convert a set of observations of possibly correlated features. PCA aims at finding the best description of the data by using a limited number of PCs. PCs are then ranked according to the explained variance in descending order(Lever et al., 2017). Given the 80% on explained variance, the raw calculation of the PCA on the features resulted in a mean of 420 Principal Components in every CV fold (range 419-421).

Step 4: Scale featurewise [from 0 to 1], zero-out completely non-finite features.

The resulting PC scores were scaled to [0, 1]

As many ML algorithms are sensitive to scale differences between features, we scaled each variable between 0 and 1 in order to remove these effects from each training sample matrix. The scaling parameters were then applied to both the inner and outer CV cycles.

In each variable evaluation step, the SVM algorithm modelled independently linear relationships between features and classification label (HC vs. FEP). In the linear kernel space, the SVM optimized a hyperplane that maximized separability between the margins based on the closest external borders of the distributions (Support Vectors) instead of using the whole sample. Based on the trained hyperplane, the algorithm then predicted subjects’ classification of the inner CV cycle by projecting its data into the learned kernel space and measuring their geometric distance to the decision boundary. This resulted in a decision value (namely, the decision score) and a predicted classification label for every participant.

In each CV and for every combination of parameters, SVM C parameter (11 Params: geometric progression of C=2-6 to 26) was regularized. The highest weighted Balanced Accuracy (BAC) was used to select the final set of SVM models across all CV partitions. These models were combined through majority voting into a final SVM decision model that predicted the group membership of the test subjects.

Permutations: To assign statistical significance to the prediction performance of our classifiers, we performed 100 random permutations between the labels (HC vs. FEP). For each permutation, we retrained all linear SVM models in the pooled repeated nested CV experiment using the respective feature subsets obtained from the observed-label analyses. The significance of the observed out-of-training BAC was calculated as the number of events where the permuted out-of-training BAC was higher or equal to the observed BAC divided by the number of permutations performed. The significance of the model was determined at α=0.05. The permutation tests allowed us to build null-distribution of classifications. The actual classification was then plotted to determine the p value and its position with respect to the null distribution (see Supplementary Figure 4).

Visualization: To visualize the average decision function, we employed the procedure previously described(Dominic B Dwyer et al., 2018), involving:

1. In PCA space, the weight vector (w) of each SVM model was projected back to voxel space. This computation was performed for every training sample on the inner cross-validation (CV1) cycle, resulting in 100 voxel-level images for a given training partition on the outer (CV2) loop.
2. The average and standard error volumes of these 100 voxel-level w images were computed.
3. For each CV2 partition, voxels with an absolute value greater than their respective standard error multiplied by 1.96 were set to one, or to zero otherwise. This thresholding procedure extracted only those voxels that reliably contributed to the average neuroanatomical decision boundary of a given CV2 partition at the 95% confidence interval.
4. The obtained binary images were summed across all CV2 partitions and divided by the number of partitions, thus forming a single map that specified every voxel’s probability of reliably contributing to the average neuroanatomical decision boundary across the entire experiment.
5. Cortical thickness visualization involved overlaying the weight probability maps onto the FreeView thickness-based template.

Reliability for each feature is defined in terms of Cross-Validation Ratio (CVR = mean(w) / standard error(w)), where w represents the normalized individual weights from SVM models generated in the repeated nested CV scheme. Normalization is performed using the Euclidean norm of w: s = w/||w||2. This metric can be regarded as a measure of how often in the repeated nested CV framework a certain feature has been chosen in order to generate decisions. This value was used to create the overlays in Figure 1 in the main manuscript.

Leave-Site Out CV scheme: This approach is similar to the *k*-fold CV scheme, but instead of randomly dividing the sample into folds, one entire site is left out. Models are trained on the remaining sites, and the predictions are applied to the left-out site. This design provides a measure of how models generalize across different geographic sites(Dominic B. Dwyer et al., 2018; Koutsouleris et al., 2018).

**4. Machine Learning pipeline**

To enable the unbiased estimation of the prediction and to prevent information leaking between subjects used for training and validating the models, we used repeated nested cross-validation (CV)12,25. Both the CV1 (inner cycle) and the CV2 (outer cycle) were randomly split in 10 folds and shuffled respectively with 10 permutations. In the inner loop of cross-validation, the training data were preprocessed using different preprocessing pipelines, as described above. Training samples were analyzed with a linear Support Vector Machine (SVM) and models, which maximized the balanced accuracy (BAC, (Sensitivity+Specificity)/2)), were selected.

**5. Association between decision scores and clinical variables by means of Support Vector Regression**

In order to maintain the multivariate framework, support vector regression (SVR) was chosen to perform regression analysis between the predictive decision scores (as variable to be predicted) obtained by the SVM analyses and clinical variables of interests (as predictors). As shown by previous literature in the field(Kambeitz-Ilankovic et al., 2019; Koutsouleris et al., 2014), SVR has the ability to generate unbiased models that generalize well across the population.

To train and cross-validate our models, a nested CV framework was applied using the open-source machine learning tool NeuroMiner (https://www.pronia.eu/neurominer/). More specifically, for the present analysis both the CV1 (inner cycle) and the CV2 (outer cycle) were randomly split in 5 folds (due to the limited number of participants in this sub-analysis) and shuffled respectively with 10 permutations. This maximized the data available to the machine learning process while both generating robust parameter estimates and avoiding any overfitting. Each training sample at the CV1 loop was first processed using feature-wise standardization: before entering the SVR, the features scores were scaled to [0, 1] and missing data were imputed using the Euclidean median of the 7 nearest neighbors. The SVR algorithm determined an optimal regression-fitting function applying optimization to C (regularization) parameter (7 parameters, C progression from 2-6 to 1) and to *ν*-parameter, which was optimized in the range of [0.05 0.075 0.1 0.125 0.15 0.2] within the CV1 cycle. CV1 test partition predictions across the *ν* range were generated and the optimal *ν* at the parameter showing the lowest average Mean Absolute Error (MAE) across the respective CV1 cycle was picked. The model trained with this *ν* parameter was then applied to the respective CV2 subject and the resulting predictions were averaged.

**Supplementary Results**

**Different preprocessing pipeline**

We built a classification modifying a step in the preprocessing pipeline. Specifically, we regressed out the covariates on the whole sample, instead of regressing them only on HC and then applying on the whole sample. This allowed a classification with a BAC of 61.8%, sensitivity 58.8%, specificity 64.7% and AUC 65%.

**Heterogeneity assessment**

First, we decided to further investigate our sample by assessing individually male and female. Male patients were better classified from male healthy individuals compared to the female groups. More precisely, the classification between male FEP and HC reached a BAC of 65.4%, with sensitivity in recognizing FEP of 77.6% and specificity of 53.1%. AUC was 69%, PPV 67.5% and PPN 65.4%. On the other hand, the classification of the female subjects reached a BAC of 62.9%, with sensitivity of 46.7%, specificity of 79.1% and an AUC of 71%. PPV and PPN were 58.3% and 70.2%, respectively.

Second, we assessed differences in the MRI scans between centers, given that five centers out of eight were equipped with a Siemens machine (namely, Jena, Prague, Verona, Rome and Basel, see Supplementary). The sample was divided accordingly.

A model was then trained on 457 subjects (195 FEP and 262 HC) scanned with a Siemens machine. The model reached a 69.5% BAC, sensitivity 63.1% and specificity 76%. AUC was 73%, and PPV and NPV 66.1% and 73.4%, respectively. This model was then tested on the remaining patients scanned with different MRI machines. This validation test had a drop in BAC to 57.3%.

We then trained a model based on the three remaining centers (Munich, Santander and London) that did not use a Siemens machine, for a total of 335 individuals (186 FEP and 149 HC). This model reached a BAC of 66%, with sensitivity 68.8%, specificity 63.1%, AUC 70%, PPV 69.9% and NPV 61.8%. When applied to the other 5 centers as test samples, the BAC dropped to 50%.

Finally, we combined the two criteria and trained the classifier with male subjects scanned on a Siemens machine and those scanned on the remaining three centers. We repeated the analysis with female subjects only. The best result was obtained combining male participants with the Siemens scan. This model trained on 233 subjects (116 FEP/117 HC) reached a BAC of 70.4%, with sensitivity of 66.7% and specificity of 73.5%. AUC was 73%, PPV 71.6% and NPV 69.4%. We tested this model on males scanned with others MRI machines (209 subjects, 130 FEP/79 HC) and the BAC dropped to 62.8%.

For comparison, we resampled our original sample by randomly removing 100 participants each time (see Table 2 for further information). We could not see any improvement in BAC, which remained between 66.8% and 62%.

To further assess the possible impact of sex and its interaction with sites or label, we performed a post-hoc multifactorial GLM analysis (SPSS was used). Specifically, we tested whether the interaction of gender, site and labels (namely, sex\*site, sex\*label, and sex\*site\*label) was associated with the prediction scores (as a target dependent variable) obtained after the classification. All the interaction resulted non-significant. More in detail, sex\*site: p=0.075 (F 1.85); , sex\*label: p=0.32 (F 0.98); sex\*site\*label: p=0.97 (F 0.25).

**Supplementary Tables**

**Supplementary Table 1.** Total sample demographics (chi square for nominal variables and t-test for continuous variables)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **FEP** | **HC** | **Sign.** |
| **Sex (F/M)**(n. 428FEP/448HC) | 147/281 | 230/218 | χ2(1)=10.2 p<0.05 |
| **Age**(n. 428FEP/448HC) | 27.14 (±7.37) | 27.06 (±6.32) | t(1,875) =0.38 p=0.702 |
| **Yrs of Education**(n. 284FEP/257HC) | 12.46 (±3.27) | 14.36 (±3.0) | t(1,540)=-6.9 p<0.05 |
| **Ethnicity (caucasian%)**(n. 261FEP/300HC) | 78% | 95% | χ2(1)=36.6 p<0.05 |
| **Handedness (r/ambi/l)**(n. 245FEP/308HC) | 216/6/23 | 285/10/13 | χ2(2)=6.2 p=0.045 |
| **Tobacco smoke**(n. 206FEP/213HC) | 68% | 29% | χ2(1)=28.8 p<0.05 |
| **Lifetime Cannabis use**(n. 180FEP/103HC) | 36% | 22.5% | χ2(1)=5.8 p<0.05 |
| **Age at onset**(n. 160FEP) | 27.15 (±8.07) |  |  |
| **DUI (month)**(n. 146FEP) | 4.8 (±8.32) |  |  |
| **N. prior hospitalizations**(n. 152FEP) | 0.8 (±0.9) |  |  |
| **Chlorpromazine equival.**(n. 253FEP) | 4347.6(±9374) |  |  |
| **PANSS pos**(n. 223FEP) | 14.33 (±6.6) |  |  |
| **PANSS neg**(n. 223FEP) | 15.31 (±7.6) |  |  |
| **PANSS gen**(n. 223FEP) | 31.25 (±12.5) |  |  |
| **PANSS tot**(n. 223FEP) | 60.83 (±22.6) |  |  |
| **GAF**(n. 256FEP) | 55.96 (±18.9) |  |  |
| **CGI**(n. 136FEP) | 3.44 (±1.38) |  |  |

CGI: clinical global impression; DUI: duration of untreated illness; FEP: first episode psychosis; GAF: global assessment of functioning; HC: healthy controls; PANSS: positive and negative symptoms scale.

**Supplementary Table 2.** Demographic characteristics per site (chi square for nominal variables and t-test for continuous variables)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Site 1** | **Site 2** | **Site 3** | **Site 4** | **Site 5** | **Site 6** |  **Site 7** | **Site 8** | **Statistics** |
| **Age** | 29.10(±5.72) | 25.92(±5.18) | 29.65(±6.53) | 27.07(±7.54) | 26.1(±5.8) | 21.86(±3.90) | 29.59(±8.54) | 26.2(±7.3) | p<0.001 |
| **Sex** | 18/31 | 66/64 | 32/31 | 89/118 | 47/73 | 24/23 | 71/77 | 29/81 | p=0.34 |
| **Age at onset** | 27.5 (±7) | NA | 30.8 (±7) | NA | NA | 20.1 (±5) | 29.7 (±9) | 26.0 (±5) | p<.001 |
| **DUI** | 7.9 (±12) | NA | 4.2 (±7) | 7.09 (±1) | NA | 3.4 (±5) | NA | 2.2 (± 2) | p=0.03 |
| **PANSS pos** | 20.5 (±8.1) | NA | 13.7 (±4.1) | 14.4 (±5.8) | NA | 15.4 (±6.8) | 15.6 (±7.1) | 13.5 (±8.6) | p=0.132 |
| **PANSS****neg** | 9.8 (±2.7) | NA | 15.0 (±4.9) | 15.6 (±6.1) | NA | 18.9 (±7.6) | 19.9 (±7.1) | 12.6 (±9.8) | p=0.003 |
| **PANSS****gen** | 37.14 (±14.7) | NA | 31.79 (±6.8) | 29.65 (±7.1) | NA | 40.12 (±10.2) | 36.03 (±10.5) | 23.7 (±19.1) | p=0.0001 |
| **GAF** | 50.8 (±33.8) | NA | 73.8 (±13.1) | 56.7 (±16.3) | 58.9 (±16.3) | 45.2 (±15.8) | 50.1 (±14.7) | NA | p=0.0001 |

DUI: duration of untreated illness; GAF: global assessment of functioning; PANSS: positive and negative symptoms scale.

**Supplementary table 3.** MR scanner systems and structural MRI sequence parameters used at the respective ClassiFEP sites

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Site | Model | Field Strength | Flip angle(°) | Coil Channel | Voxel Size (mm) | TR (ms) | TE (ms) | FOV | Slices |
| Jena | Siemens TrioTim | 3T | 9 |  | 1.0 x 1.0 x 1.0 | 2300 | 3.03 | 256x256 | 192 |
| Prague | Siemens TrioTim | 3T | 10 | 12 | 1.0 x 1.0 x 1.0 | 2300 | 4.63 | 256x256 |  |
| Santander | Philips Achieva | 3T | 8 |  | 1.0 x 1.0 x 1.0 | 3000 | 3.9 | 256x256 |  |
| Verona | Siemens Magnetom Allegra | 3T | 15 |  | 1.0 x 1.0 x 1.0 | 2060 | 3.93 | 256x256 |  |
| Rome | Siemens Allegra RM | 3T | 15 |  | 1.0 x 1.0 x 1.0 | 2400 | 7.9 | 256x256 |  |
| London | General Electric Signa HDx scanner | 3T | 20 |  | 1.09 × 1.09 × 1.1 | 6980 | 2.81 | 256 × 256 | 196 |
| Basel | Siemens Magnetom Verio | 3T | 8 | 12 | 1.0 x 1.0 x 1.0 | 2000 | 3.4 | 256x256 | 176 |
| Munich | Philips Ingenia | 3T | 8 | 32 | 0.97 x 0.97 x 1.0 | 9500 | 5.5 | 250x250 | 190 |

**Supplementary Table 4.** Classification analyses by site

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Site** | **BAC** | **AUC** | **Sens** | **Spec** | **PPV** | **NPV** |
| Site1 | 57 | 0.56 | 34.8 | 79.2 | 61.5 | 55.9 |
| Site 2 | 54.5 | 0.63 | 12.9 | 94 | 4 | 77.7 |
| Site 3 | 54 | 0.56 | 63.6 | 42.3 | 55.3 | 53.2 |
| Site 4 | 66.7 | 0.72 | 60.9 | 73.6 | 65.1 | 69.6 |
| Site 5 | 64.5 | 0.67 | 72 | 58.1 | 73 | 55.8 |
| Site 6 | 65.1 | 0.69 | 62.8 | 66.2 | 68.6 | 60.5 |
| Site 7 | 60 | 0.66 | 33.3 | 86.8 | 57.1 | 71.1 |
| Site 8 | 64 | 0.71 | 70.7 | 57.3 | 67.8 | 60.8 |

AUC: Area Under the Curve; BAC: Balanced Accuracy; NPV: Negative Predictive Value; PPV: Positive Predictive Value; Sens: sensitivity; Spec: specificity

**Supplementary Figures**

**Supplementary Figure 1**. Distribution of the Euler Number across our sample and visual correlation with reconstruction of FreeSurfer images. As shown, the lower the Euler Number, the worse the reconstruction.



**Supplementary Figure 2.** Correlation between Decision Scores and Euler Number in the model built on all the images (without excluding images labeled as „poor reconstruction“).



**Supplementary Figure 3.** Classification gain after the main analysis. Comapred to a null-model, the classification improved the ability to distinguish between the group by a 16.9% for FEP and 15.6% for HC. These percentage are a measure of the size of the improvement obtained with the SVM.

****

**Supplementary figure 4.** Calculation of p value. We randomly permuted 100 times the labels and calculated a null distribution on the classification models based on the permutation analysis. The red line indicates a significant BAC with a p value < 0.01.

****

**Supplementary Figure 5.** The reliability of predictive voxels on the Leave-Site Out analysis was plotted via a cross-validation ratio map, using FreeView (version 5.3 http://surfer.nmr.mgh.harvard.edu/) to overlay the maps on the single-participant template. A threshold of ±2, which corresponded to an α-level of .05 was set, as presented by Koutsouleris et al.(Koutsouleris et al., 2018). In this way, only significant voxels after CV are presented. The color scale indicates increased *vs*. decreased thickness in individuals with FEP compared with HC. The cool color scale indicates reduced thickness and the warm color scale increased thickness in FEP compared to HC. Subpanels 1a and 1b show the external surface of the brain, with cluster of voxels in the superior temporal and planum temporale. Subpanels 2a and 2b show the ventral surface of the brain, while subpanels 3a and 3b show the superior and inferior surface of the brain, where so significant vowels are present.



**References**

Dwyer, Dominic B, Cabral, C., Kambeitz-Ilankovic, L., Sanfelici, R., Kambeitz, J., Calhoun, V., Falkai, P., Pantelis, C., Meisenzahl, E., Koutsouleris, N., 2018. Brain Subtyping Enhances The Neuroanatomical Discrimination of Schizophrenia. Schizophr. Bull. 44, 1060–1069. https://doi.org/10.1093/schbul/sby008

Dwyer, Dominic B., Falkai, P., Koutsouleris, N., 2018. Machine Learning Approaches for Clinical Psychology and Psychiatry. Annu. Rev. Clin. Psychol. 14, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. &quot;Mini-mental state&quot;. A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12, 189–98.

Gorgens, K.A., 2011. Structured Clinical Interview For DSM-IV (SCID-I/SCID-II), in: Encyclopedia of Clinical Neuropsychology. Springer New York, New York, NY, pp. 2410–2417. https://doi.org/10.1007/978-0-387-79948-3\_2011

Hansen, L.K., Larsen, J., Nielsen, F.A., Strother, S.C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., Paulson, O.B., 1999. Generalizable patterns in neuroimaging: how many principal components? Neuroimage 9, 534–44. https://doi.org/10.1006/nimg.1998.0425

Jablensky, A., Sartorius, N., Ernberg, G., Anker, M., Korten, A., Cooper, J.E., Day, R., Bertelsen, A., 1992. Schizophrenia: manifestations, incidence and course in different cultures. A World Health Organization ten-country study. Psychol. Med. Monogr. Suppl. 20, 1–97.

Kambeitz-Ilankovic, L., Haas, S.S., Meisenzahl, E., Dwyer, D.B., Weiske, J., Peters, H., Möller, H.-J., Falkai, P., Koutsouleris, N., 2019. Neurocognitive and neuroanatomical maturation in the clinical high-risk states for psychosis: A pattern recognition study. NeuroImage. Clin. 21, 101624. https://doi.org/10.1016/j.nicl.2018.101624

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., Reiser, M., Pantelis, C., Meisenzahl, E., 2014. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophr. Bull. 40, 1140–53. https://doi.org/10.1093/schbul/sbt142

Koutsouleris, N., Kahn, R.S., Chekroud, A.M., Leucht, S., Falkai, P., Wobrock, T., Derks, E.M., Fleischhacker, W.W., Hasan, A., 2016. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. The Lancet Psychiatry 3, 935–946. https://doi.org/10.1016/S2215-0366(16)30171-7

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D.B., Paolini, M., Chisholm, K., Kambeitz, J., Haidl, T., Schmidt, A., Gillam, J., Schultze-Lutter, F., Falkai, P., Reiser, M., Riecher-Rössler, A., Upthegrove, R., Hietala, J., Salokangas, R.K.R., Pantelis, C., Meisenzahl, E., Wood, S.J., Beque, D., Brambilla, P., Borgwardt, S., PRONIA Consortium, 2018. Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression. JAMA Psychiatry 75, 1156. https://doi.org/10.1001/jamapsychiatry.2018.2165

Lasalvia, A., Tosato, S., Brambilla, P., Bertani, M., Bonetto, C., Cristofalo, D., Bissoli, S., De Santi, K., Lazzarotto, L., Zanatta, G., Marrella, G., Mazzoncini, R., Zanoni, M., Garzotto, N., Dolce, C., Nicolau, S., Ramon, L., Perlini, C., Rambaldelli, G., Bellani, M., Tansella, M., Ruggeri, M., 2012. Psychosis Incident Cohort Outcome Study (PICOS). A multisite study of clinical, social and biological characteristics, patterns of care and predictors of outcome in first-episode psychosis. Background, methodology and overview of the patient sample. Epidemiol. Psychiatr. Sci. 21, 281–303. https://doi.org/10.1017/S2045796012000315

Lever, J., Krzywinski, M., Altman, N., 2017. Principal component analysis. Nat. Methods 14, 641–642. https://doi.org/10.1038/nmeth.4346

Magni, E., Binetti, G., Bianchetti, A., Rozzini, R., Trabucchi, M., 1996. Mini-Mental State Examination: a normative study in Italian elderly population. Eur. J. Neurol. 3, 198–202. https://doi.org/10.1111/j.1468-1331.1996.tb00423.x

Rosen, A.F.G., Roalf, D.R., Ruparel, K., Blake, J., Seelaus, K., Villa, L.P., Ciric, R., Cook, P.A., Davatzikos, C., Elliott, M.A., Garcia de La Garza, A., Gennatas, E.D., Quarmley, M., Schmitt, J.E., Shinohara, R.T., Tisdall, M.D., Craddock, R.C., Gur, R.E., Gur, R.C., Satterthwaite, T.D., 2018. Quantitative assessment of structural image quality. Neuroimage 169, 407–418. https://doi.org/10.1016/j.neuroimage.2017.12.059

Ruschhaupt, M., Huber, W., Poustka, A., Mansmann, U., 2004. A compendium to ensure computational reproducibility in high-dimensional classification tasks. Stat. Appl. Genet. Mol. Biol. 3, Article37. https://doi.org/10.2202/1544-6115.1078

Savalia, N.K., Agres, P.F., Chan, M.Y., Feczko, E.J., Kennedy, K.M., Wig, G.S., 2017. Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. Hum. Brain Mapp. 38, 472–492. https://doi.org/10.1002/hbm.23397

Schnack, H.G., Kahn, R.S., 2016. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. Front. psychiatry 7, 50. https://doi.org/10.3389/fpsyt.2016.00050

Sheehan, D. V, Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry 59 Suppl 20, 22-33;quiz 34-57.

Tordesillas-Gutierrez, D., Ayesa-Arriola, R., Delgado-Alvarado, M., Robinson, J.L., Lopez-Morinigo, J., Pujol, J., Dominguez-Ballesteros, M.E., David, A.S., Crespo-Facorro, B., 2018. The right occipital lobe and poor insight in first-episode psychosis. PLoS One 13, e0197715. https://doi.org/10.1371/journal.pone.0197715