Supplementary material

Number of Supplementary Figures/Tables: 18/9

Two Distinct Neuroanatomical Subtypes of Schizophrenia Revealed Using Machine Learning

SI Methods

Study sample

Subjects for this consortium-based study were collected from previous studies conducted in three different sites (USA, Germany and China) (Satterthwaite *et al.*, 2010; Wolf *et al.*, 2014; Zhang *et al.*, 2015; Zhu *et al.*, 2016; Zhuo *et al.*, 2016). Parts of these samples were also formerly used for a multisite standard case-control study by our group (Rozycki *et al.*, 2018).

Site 1 (USA): This study was approved by the Institutional Review Board of the University of Pennsylvania. Written informed consent was obtained from each participant. Recruitment and assessment were performed by trained clinical investigators. Diagnostic assessment utilized the Structured Clinical Interview for DSM-IV (SCID). Subjects were not enrolled if they had a history of substance abuse or dependence (excluding nicotine) in the past six months or a positive urine drug screen on the day of the study. Healthy control (HC) subjects were excluded if they met criteria for any DSM-IV psychiatric disorder. For patient samples, the Scale for the Assessment of Positive Symptoms (SAPS) (Andreasen, 1984) and the Scale for the Assessment of Negative Symptoms (SANS) (Andreasen, 1983) were used.

Site 2 (Germany): Subjects were recruited at the Department of Psychiatry and Psychotherapy at Ludwig-Maximilians University, Munich, Germany. The study protocols were approved by the ethics committee of Ludwig-Maximilians University. Subjects provided their written informed consent prior to MRI and clinical examination. Patient recruitment and assessment was performed by trained clinical investigators. Assessment included the SCID for Axis I & II disorders (SCID-I/-II), a semi-standardized clinical interview for the assessment of medical and psychiatric history, review of medical records and psychotropic medications, and the evaluation of disease severity and psychopathology by means of the Positive and Negative Syndrome Scale (PANSS) (Kay *et al.*, 1987). Patients received a consensus diagnosis by two experienced psychiatrists at study inclusion. Participants were excluded if they had other psychiatric and/or neurological diseases, past or present regular alcohol abuse, and/or consumption of illicit drugs, past head trauma with loss of consciousness or electroconvulsive treatment, insufficient knowledge of German, IQ < 70, and age < 18 or > 65 years.

Site 3 (China): The study was approved by the Ethics Committee of Tianjin Medical University General Hospital, Tianjin, China. Written informed consent was obtained from each subject before study enrollment. Diagnosis of schizophrenia was determined based on the consensus of two expert clinical psychiatrists using DSM-IV (SCID). Subject inclusion criteria were age (16–60 years) and right-handedness. Subject exclusion criteria were MRI contraindications, pregnancy, and histories of systemic medical illness, central nervous system disorder and head trauma, and substance abuse within the last 3 months or lifetime history of substance abuse or dependence. For HC, the additional exclusion criteria were a history of psychiatric disease and first-degree relatives with a psychotic disorder. PANSS scores were assessed for schizophrenia patients.

Education Level

Educational attainment was coded as follows: ordinal scale of 1 for education up to 12 years of age, 2 for education up to 16 years of age, and 3 for education up to 18 years of age. Education variables were not available in Site 3 (China).

Positive symptoms and negative symptoms

Disease severity and psychopathology were evaluated by means of the PANSS(Kay *et al.*, 1987) in Site 2 (Germany) and Site 3 (China) while SAPS (Andreasen, 1984) and SANS (Andreasen, 1983, 1989) were assessed in Site 1 (USA). These symptom rating scales have been widely used in schizophrenia research. PANSS consists of PANSS-positive, PANSS-negative and PANSS-general sub-domains (van Erp *et al.*, 2014). For consistency of the symptom scales across sites, the SAPS and SANS were converted into the PANSS-positive and PANSS-negative using an established method (van Erp *et al.*, 2014).

Image acquisition

In Site 1 (USA), images were acquired at a 3T TIM TRIO scanner (Siemens, Erlangen, Germany) using T1-weighted 3D magnetization-prepared rapid acquisition with gradient echo sequences (MPRAGE) (TR = 1810 ms, TE= 3.51 ms, TI = 1100 ms, flip angle = 9 degree, FOV = 240 mm x 180 mm, matrix = 256 × 192, slices = 160, slice/skip thickness = 1 mm/0 mm).

In Site 2 (Germany), T1-weighted MPRAGE (TR = 11.6 ms, TE = 4.9 ms, FOV = 230 mm, matrix = 512 x 512, 126 contiguous axial slices of 1.5 mm thickness, voxel size = 0.45 x 0.45 x 1.5 mm) were acquired at a 1.5 T Magnetom Vision scanner (Siemens, Erlangen, Germany).

In Site 3 (China), images were acquired at a 3T MR system (Discovery MR750, General Electric, Milwaukee, WI, USA). Sagittal 3D T1-weighted images were acquired using a brain volume sequence (BRAVO) (TR = 8.2 ms, TE = 3.2 ms, TI = 450 ms, flip angle = 12 degree, FOV = 256 mm x 256 mm, matrix = 256 x 256, slice thickness = 1 mm, no gap, 188 sagittal slices).

Image preprocessing

A set of extensive quality assurance procedures were applied using both manual verification and automated flags. Raw T1-images were manually examined for motion, image artifacts, or restricted field-of-view. Images were corrected for magnetic field inhomogeneity (Tustison *et al.*, 2010) and a multi-atlas, multi-warp segmentation method (MUSE) (Doshi *et al.*, 2016) was used to segment each individual's images into anatomical regions of interest (ROIs) consisting of gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF). The voxel-wise regional volumetric maps (Davatzikos *et al.*, 2001) were generated for GM, WM and CSF tissues by registering skull-stripped T1-images to a template residing in the MNI-space using a deformable registration method (Ou *et al.*, 2011). The processed images were also manually evaluated (authors D. S. and G. E.) for pipeline failures, such as for poor brain extraction, poor tissue segmentation, and registration errors. Furthermore, automated procedures flagged images based on outlying values of quantified metrics (i.e., regional volumes) and those flagged images were re-evaluated.

HYDRA

In HYDRA (Varol *et al.*, 2017), classification is performed through the separation of healthy controls from patients by a convex polytope formed by linear maximum-margin classifiers. Subtyping is carried out by clustering patients through their association with different faces of the polytope referred to as hyperplanes. HYDRA consists of the following main steps: an initialization followed by iterations of assignment and polytope solutions, and the consensus of clustering results. Specifically, HYDRA initializes the assignments of patients into clusters by sampling K unit length hyperplanes obtained considering the space of all pairwise differences between patients and HC. The K unique hyperplanes are chosen by employing determinantal point processes (DPP) (Kulesza and Taskar, 2012), a sampling technique that samples diverse directions of disease. The sampled hyperplanes are subsequently used to estimate the initial clustering assignments (S^-). As the estimated solution may vary depending on the initialization, a multi-initialization strategy is implemented by the DPP. The final clustering results are achieved based on a consensus of clustering solutions. The HYDRA algorithm (Varol *et al.*, 2017) is summarized as:

Input: $X \in R^{n \times d}$, $Y \in \{-1, +1\}^n$ (training signals with n-subjects, d-imaging features), K (number of subtypes or hyperplanes)

Output: $W \in R^{d \times K}$, $b \in R^{1 \times K}$ (classifier); S^- (clustering assignment)

Initialization: Initialize S^-

Loop: Repeat until convergence (or a fixed number of iterations)

```
    Fix $S^-$, solve for W and b
    Fix W and b, solve for $S^-$
```

HYDRA analyses were carried out using the following parameters: 50 iterations between estimating hyperplanes and cluster estimation, 20 clustering consensus steps, 0.25 regularization parameter and 10 cross-validation folds. The clustering performance of HYDRA was assessed by taking into account the stability of the obtained solutions. The adjusted Rand index (Hubert and Arabie, 1985) was used to quantify the similarity between clustering results in a 10-folds cross-validated fashion by taking into account the clustering stability between folds. Hence, the ARI calculates how consistently common subjects are placed in the same clusters despite variations in the sample composition across folds. The ARI corrects for grouping by chance, providing a more conservative estimation of the overlap. An ARI value equal to 1 indicates a perfect clustering.

MIDAS

MIDAS is a recently published and validated method for voxel-based group comparisons (Varol *et al.*, 2018). It overcomes limitations of commonly used voxel-based analysis due to ad hoc filtering of image maps by arbitrary and fixed Gaussian filters prior to applying a general linear model. MIDAS effectively determines the regionally varying, anisotropic filtering of any image data that optimally captures group differences. Voxel-wise regional volumetric maps were compared between the groups using MIDAS. MIDAS parses any set of image maps (volumetric maps, herein), using a sufficiently large set of overlapping neighborhoods (P), and performs regional discriminative analysis based on least squares support vector machines (LS-SVM). LS-SVM aims to relate the imaging features $X \in R^{n \text{ x } v}$ (n-subjects and v-dimensional imaging features) with group variable $Y \in R^n$ via a weight vector (w). The regional pattern that best discriminates between groups is equivalent to filtering locally by an optimal kernel whose coefficients are the weights of the discriminant. A statistic for a particular voxel is then computed by modulating the total contribution of each voxel to the estimated local activation patterns (a) with the total predictive power of the respective machine learners (Varol *et al.*, 2018).

$$s_i = \frac{\sum_{p=1}^{P} a_i^p}{\sum_{p=1}^{P} \| w^p \|_2^2}$$

where, $a \propto \frac{1}{n}(X - \bar{X})^T(X - \bar{X})w,$ and $\sum_{p=1}^{P} \parallel w^p \parallel_2^2$ is the sum of the inverse predictive power of all learners, in which voxel i participates. This voxel-wise statistic indicates the degree of participation of this voxel in all partially overlapping regional filters that contain that voxel. Finally, the p-value corresponding to the voxel statistic is analytically obtained by approximating permutation tests.

In MIDAS, the voxel-wise regional volumetric maps (Davatzikos *et al.*, 2001) of GM, WM or CSF with dimensions 182 x 218 x 182 were used to assess the voxel-wise neuroanatomical differences between the groups. MIDAS analyses were carried out using the following parameters: 15 neighborhood radius in voxels, 500 neighborhoods, and 0.1 regularization parameter. The voxel-wise statistical significance values (p-values) of

MIDAS were further corrected for multiple comparison by false discovery rate (FDR) (FDR-p < 0.05) and then used as a mask to show the effect size (Cohen's d) maps between the groups.

Permutation tests for the subtypes

Permutation testing is a well-known framework, which is extensively used when the underlying null distribution is unknown or hard to estimate (Nichols and Homes, 2001). To examine the null distribution of the subtype stability, subtyping analyses were carried out in HC samples, where disease-related variability is not present. For this, the HC samples (n = 364) were randomly assigned to a HC group (~20%) and a pseudo-patient group (~80%), and HYDRA analysis was performed. These samples were permuted 50 times, and HYDRA was run each time. To fairly compare these results with the clustering results obtained using the actual patient group, we selected analogously-sized HC and patient groups (~20% and ~80% of 364, respectively) so that equal numbers were used as in the permutation tests. This was done so that the null distribution and the actual experiments were derived using the exact same sample sizes. Finally, the ARIs obtained in the actual experiment were compared with the null distribution of ARIs obtained in the random permutation experiments, in order to determine statistical significance.

Split-sample reproducibility

In order to investigate the reproducibility of schizophrenia subtypes, we implemented a split-sample analysis. This strategy has been widely used in the clustering literature (Ben-Hur *et al.*, 2002; Lange *et al.*, 2004). The HC samples and patient samples were divided into two halves and then HYDRA was applied in Split 1 and Split 2, independently. Voxel-wise volumetric profiles were further compared between the splits.

Leave-one-site-out validation

The main clustering results were further cross-validated using a leave-one-site-out (LOSO) method (Arlot and Celisse, 2010). In this method, HYDRA models were trained in the two data sites and then the trained models were tested in the remaining one site to identify the subtype labels (Subtype 1 or Subtype 2). This procedure was repeated for all three possible combinations of sites, as shown schematically (Figure S6). LOSO-predicted Subtype 1 and Subtype 2 assignments from all three sites were compared with the original assignments obtained by taking all the sites together. The voxel-wise GM regional patterns between LOSO-predicted Subtype 1 and Subtype 2, as well as between each subtype and HC, were evaluated.

Prevalence of the two subtypes

In our main clustering results (Figure 1), the number of schizophrenia participants in Subtype 1 (n = 192) and Subtype 2 (n = 115) turned out to be different. To ensure that distinct volumetric profiles were not influenced by variations in sample size, which affects whether or not a given effect size is statistically significant, we evaluated GM volumetric patterns of schizophrenia Subtype 1 compared with HC by randomly selecting a part of Subtype 1 samples equal in number to the number of samples in Subtype 2 (n = 115).

Reproducibility of the subtypes within sexes

Our overall sample consisted of ~40% female and ~60% male subjects. Although we applied a linear model to adjust for sex in all of the results, we further investigated the volumetric profiles of the two most reproducible subtypes in males and females separately, to ensure that our findings were not confounded by sex differences.

Table S1: List of brain regions used as features in HYDRA
(L: Left hemisphere; R: Right hemisphere; WM: White matter)

Brain regions (1-25)	**Brain regions (26-50)**	**Brain regions (51-75)**
3rd ventricle	Ventral diencephalon (R)	Anterior insula (L)
4th ventricle	Ventral diencephalon (L)	Anterior orbital gyrus (R)
Accumbens area (R)	Cerebellar vermal lobules I-V	Anterior orbital gyrus (L)
Accumbens area (L)	Cerebellar vermal lobules VI-VII	Angular gyrus (R)
Amygdala (R)	Cerebellar vermal lobules VIII-X	Angular gyrus (L)
Amygdala (L)	Basal forebrain (R)	Calcarine cortex (R)
Brain Stem	Basal forebrain (L)	Calcarine cortex (L)
Caudate (R)	Frontal lobe WM (R)	Central operculum (R)
Caudate (L)	Frontal lobe WM (L)	Central operculum (L)
Cerebellum exterior (R)	Occipital lobe WM (R)	Cuneus (R)
Cerebellum exterior (L)	Occipital lobe WM (R)	Cuneus (L)
Cerebellum WM (R)	Parietal lobe WM (R)	Entorhinal area (R)
Cerebellum WM (L)	Parietal lobe WM (L)	Entorhinal area (L)
Hippocampus (R)	Temporal lobe WM (R)	Frontal operculum (R)
Hippocampus (L)	Temporal lobe WM (L)	Frontal operculum (L)
Inferior lateral ventricle (R)	Fornix (R)	Frontal pole (R)
Inferior lateral ventricle (L)	Fornix (L)	Frontal pole (L)
Lateral ventricle (R)	Anterior limb of internal capsule (R)	Fusiform gyrus (R)
Lateral ventricle (L)	Anterior limb of internal capsule (L)	Fusiform gyrus (L)
Pallidum (R)	Posterior limb of internal capsule including cerebral peduncle (R)	Gyrus rectus (R)
Pallidum (L)	Posterior limb of internal capsule including cerebral peduncle (L)	Gyrus rectus (L)
Putamen (R)	Corpus callosum	Inferior occipital gyrus (R)
Putamen (L)	Anterior cingulate gyrus (R)	Inferior occipital gyrus (L)
Thalamus proper (R)	Anterior cingulate gyrus (L)	Inferior temporal gyrus (R)
Thalamus proper (L)	Anterior insula (R)	Inferior temporal gyrus (L)
Brain regions (76-100)	**Brain regions (101-125)**	**Brain regions (126-145)**
Lingual gyrus (R)	Occipital fusiform gyrus (L)	Subcallosal area (R)
Lingual gyrus (L)	Opercular part of inferior frontal gyrus (R)	Subcallosal area (L)
Lateral orbital gyrus (R)	Opercular part of inferior frontal gyrus (L)	Superior frontal gyrus (R)
Lateral orbital gyrus (L)	Orbital part of inferior frontal gyrus (R)	Superior frontal gyrus (L)
Middle cingulate gyrus (R)	Orbital part of inferior frontal gyrus (L)	Supplementary motor cortex (R)
Middle cingulate gyrus (L)	Posterior cingulate gyrus (R)	Supplementary motor cortex (L)
Medial frontal cortex (R)	Posterior cingulate gyrus (L)	Supramarginal gyrus (R)
Medial frontal cortex (L)	Precuneus (R)	Supramarginal gyrus (L)
Middle frontal gyrus (R)	Precuneus (L)	Superior occipital gyrus (R)
Middle frontal gyrus (L)	Parahippocampal gyrus (R)	Superior occipital gyrus (L)
Middle occipital gyrus (R)	Parahippocampal gyrus (L)	Superior parietal lobule (R)
Middle occipital gyrus (L)	Posterior insula (R)	Superior parietal lobule (L)
Medial orbital gyrus (R)	Posterior insula (L)	Superior temporal gyrus (R)
Medial orbital gyrus (L)	Parietal operculum (R)	Superior temporal gyrus (L)
Postcentral gyrus medial segment (R)	Parietal operculum (L)	Temporal pole (R)
Postcentral gyrus medial segment (L)	Postcentral gyrus (R)	Temporal pole (L)
Precentral gyrus medial segment (R)	Postcentral gyrus (L)	Triangular part of the inferior frontal gyrus (R)
Precentral gyrus medial segment (L)	Posterior orbital gyrus (R)	Triangular part of the inferior frontal gyrus (L)
Superior frontal gyrus medial segment (R)	Posterior orbital gyrus (L)	Transverse temporal gyrus (R)
Superior frontal gyrus medial segment (L)	Planum polare (R)	Transverse temporal gyrus (L)
Middle temporal gyrus (R)	Planum polare (L)	
Middle temporal gyrus (L)	Precentral gyrus (R)	
Occipital pole (R)	Precentral gyrus (L)	
Occipital pole (L)	Planum temporale (R)	
Occipital fusiform gyrus (R)	Planum temporale (L)	

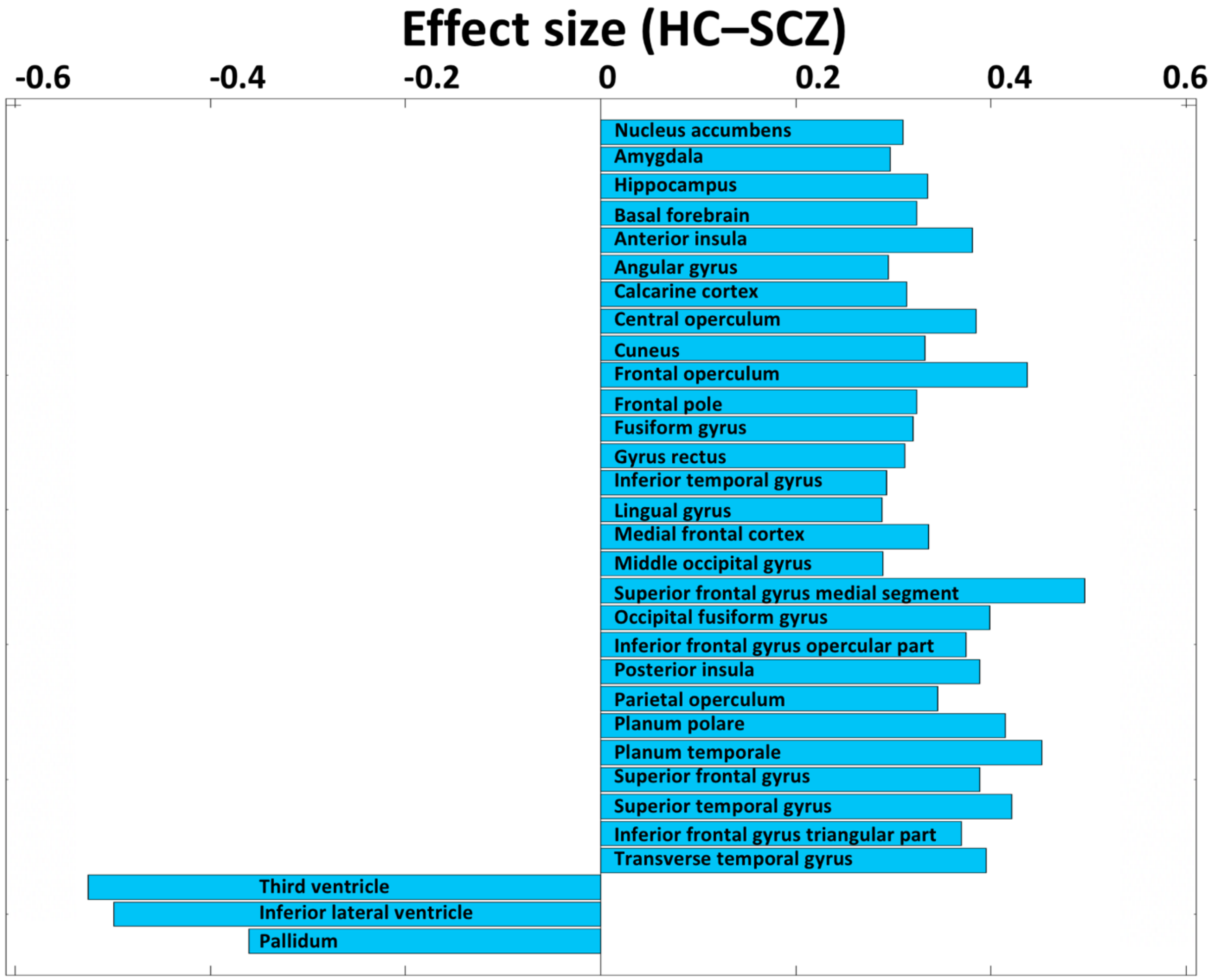

Figure S1: Key regional volume differences between healthy controls (HC) (n = 364) and schizophrenia (SCZ) (n = 307): ROIs with the highest effect size (absolute effect size > 0.28 and FDR-p < 0.05) are displayed. Note that both volume decreases and increases are observed in this standard case-control comparison.

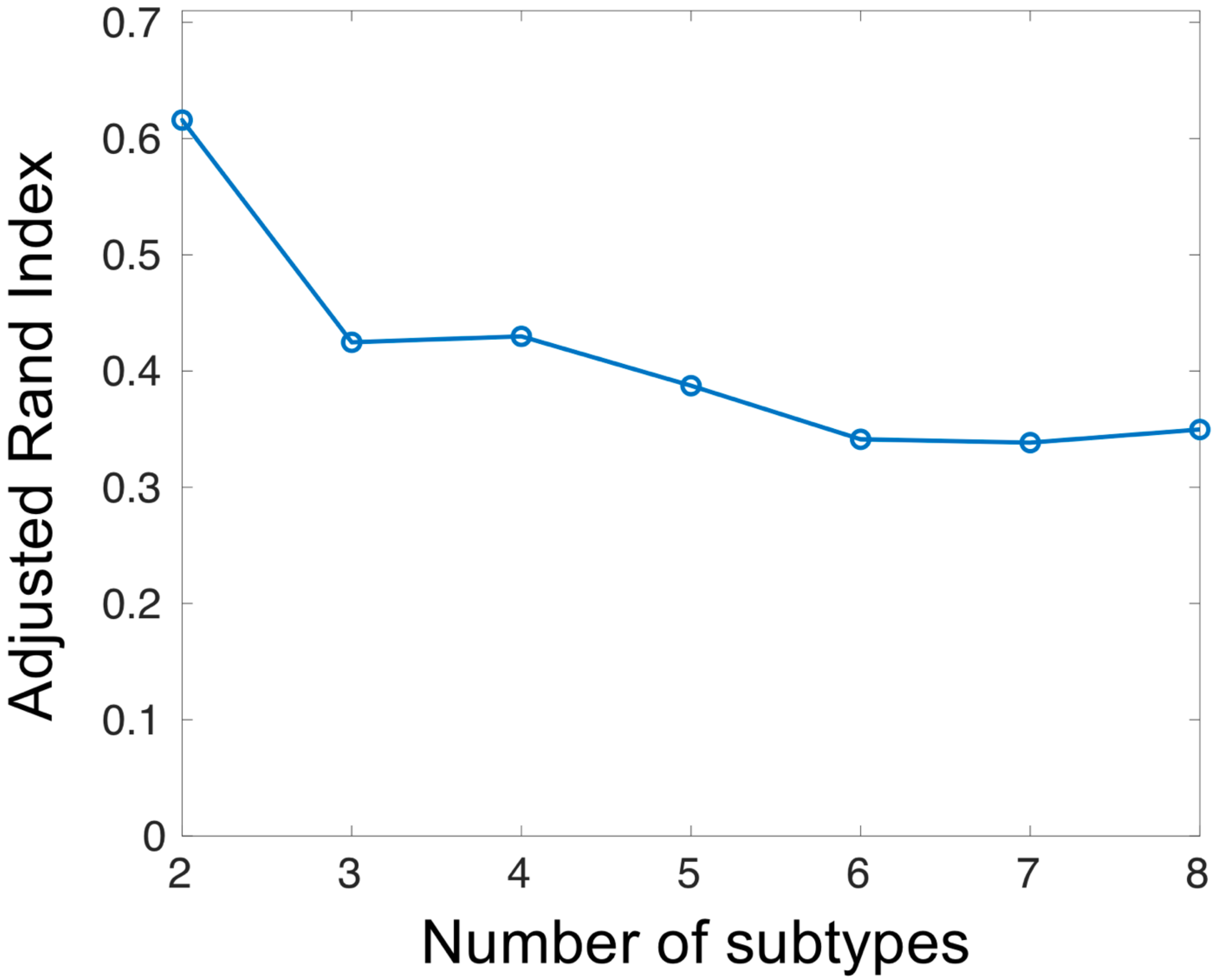

Figure S2: Cross-validated stability of schizophrenia subtypes: Adjusted Rand Index (ARI) vs. number of subtypes (K) indicating high reproducibility for K = 2.

Table S2 SCZ sample number distribution across the subtypes and data sites

SCZ subtypes	Total (n = 307)	Site 1 (n = 96)	Site 2 (n = 145)	Site 3 (n = 66)
K = 2	192	64	91	37
	115	32	54	29
K = 3	147	63	47	37
	87	27	32	28
	73	6	66	1
K = 4	71	7	61	3
	50	18	20	12
	111	46	36	29
	75	25	28	22
K = 5	52	24	8	20
	48	15	19	14
	92	39	30	23
	67	3	64	0
	48	15	24	9
K = 6	41	15	17	9
	68	27	24	17
	44	14	19	11
	42	16	18	8
	58	1	57	0
	54	23	10	21
K = 7	34	11	15	8
	40	17	9	14
	51	5	44	2
	50	14	27	9
	60	23	19	18
	33	10	18	5
	39	16	13	10
K = 8	37	12	21	4
	38	16	8	14
	29	4	18	7
	38	15	13	10
	43	18	10	15
	46	10	31	5
	35	0	33	2
	41	21	11	9

Table S3: Demographic comparison between K = 2 subtypes (SCZ1 and SCZ2)

	Age	Sex
SCZ1 vs. SCZ2 p-value	0.109	0.910

Table S4: Demographic comparison among K = 3 (SCZ1, SCZ2 and SCZ3)

	Age	Sex
SCZ1 vs. SCZ2 p-value	0.178	0.898
SCZ1 vs. SCZ3 p-value	0.679	0.037
SCZ2 vs. SCZ3 p-value	0.124	0.073

Table S5: Demographic comparison among K = 4 (SCZ1, SCZ2, SCZ3 and SCZ4)

	Age	Sex
SCZ1 vs. SCZ2 p-value	0.166	0.189
SCZ1 vs. SCZ3 p-value	0.327	0.349
SCZ1 vs. SCZ4 p-value	0.084	0.029
SCZ2 vs. SCZ3 p-value	0.024	0.565
SCZ2 vs. SCZ4 p-value	0.006	0.505
SCZ3 vs. SCZ4 p-value	0.378	0.141

Table S6: Demographic comparison among K = 5 (SCZ1, SCZ2, SCZ3, SCZ4 and SCZ5)

	Age	Sex
SCZ1 vs. SCZ2 p-value	0.439	0.007
SCZ1 vs. SCZ3 p-value	0.156	0.067
SCZ1 vs. SCZ4 p-value	0.026	0.0002
SCZ1 vs. SCZ5 p-value	0.370	0.101
SCZ2 vs. SCZ3 p-value	0.650	0.195
SCZ2 vs. SCZ4 p-value	0.275	0.440
SCZ2 vs. SCZ5 p-value	0.933	0.275
SCZ3 vs. SCZ4 p-value	0.447	0.021
SCZ3 vs. SCZ5 p-value	0.918	0.949
SCZ4 vs. SCZ5 p-value	0.308	0.050

Table S7: Demographic comparison among K = 6 (SCZ1, SCZ2, SCZ3, SCZ4, SCZ5 and SCZ6)

	Age	Sex
SCZ1 vs. SCZ2 p-value	0.104	0.460
SCZ1 vs. SCZ3 p-value	0.415	0.169
SCZ1 vs. SCZ4 p-value	0.127	0.323
SCZ1 vs. SCZ5 p-value	0.391	0.013
SCZ1 vs. SCZ6 p-value	0.099	0.681
SCZ2 vs. SCZ3 p-value	0.520	0.431
SCZ2 vs. SCZ4 p-value	0.775	0.715
SCZ2 vs. SCZ5 p-value	0.390	0.048
SCZ2 vs. SCZ6 p-value	0.932	0.206
SCZ3 vs. SCZ4 p-value	0.455	0.705
SCZ3 vs. SCZ5 p-value	0.927	0.303
SCZ3 vs. SCZ6 p-value	0.487	0.061
SCZ4 vs. SCZ5 p-value	0.338	0.155
SCZ4 vs. SCZ6 p-value	0.833	0.144
SCZ5 vs. SCZ6 p-value	0.358	0.002

Table S8: Demographic comparison among K = 7 (SCZ1, SCZ2, SCZ3, SCZ4, SCZ5, SCZ6 and SCZ7)

	Age	Sex
SCZ1 vs. SCZ2 p-value	0.442	0.023
SCZ1 vs. SCZ3 p-value	0.715	0.919
SCZ1 vs. SCZ4 p-value	0.814	0.797
SCZ1 vs. SCZ5 p-value	0.501	0.104
SCZ1 vs. SCZ6 p-value	0.906	0.539
SCZ1 vs. SCZ7 p-value	0.582	0.191
SCZ2 vs. SCZ3 p-value	0.585	0.008
SCZ2 vs. SCZ4 p-value	0.488	0.005
SCZ2 vs. SCZ5 p-value	0.832	0.368
SCZ2 vs. SCZ6 p-value	0.513	0.101
SCZ2 vs. SCZ7 p-value	0.158	0.307
SCZ3 vs. SCZ4 p-value	0.871	0.862
SCZ3 vs. SCZ5 p-value	0.708	0.049
SCZ3 vs. SCZ6 p-value	0.822	0.437
SCZ3 vs. SCZ7 p-value	0.287	0.118
SCZ4 vs. SCZ5 p-value	0.591	0.034
SCZ4 vs. SCZ6 p-value	0.926	0.353
SCZ4 vs. SCZ7 p-value	0.354	0.086
SCZ5 vs. SCZ6 p-value	0.593	0.346
SCZ5 vs. SCZ7 p-value	0.161	0.820
SCZ6 vs. SCZ7 p-value	0.495	0.502

Table S9: Demographic comparison among K = 8 (SCZ1, SCZ2, SCZ3, SCZ4, SCZ5, SCZ6, SCZ7 and SCZ8)

	Age	Sex
SCZ1 vs. SCZ2 p-value	0.344	0.025
SCZ1 vs. SCZ3 p-value	0.367	0.240
SCZ1 vs. SCZ4 p-value	0.463	0.514
SCZ1 vs. SCZ5 p-value	0.217	0.082
SCZ1 vs. SCZ6 p-value	0.497	0.625
SCZ1 vs. SCZ7 p-value	0.735	0.056
SCZ1 vs. SCZ8 p-value	0.622	0.389
SCZ2 vs. SCZ3 p-value	0.049	0.002
SCZ2 vs. SCZ4 p-value	0.087	0.107
SCZ2 vs. SCZ5 p-value	0.778	0.564
SCZ2 vs. SCZ6 p-value	0.746	0.059
SCZ2 vs. SCZ7 p-value	0.159	0.00008
SCZ2 vs. SCZ8 p-value	0.094	0.148
SCZ3 vs. SCZ4 p-value	0.874	0.078
SCZ3 vs. SCZ5 p-value	0.022	0.006
SCZ3 vs. SCZ6 p-value	0.091	0.099
SCZ3 vs. SCZ7 p-value	0.525	0.817
SCZ3 vs. SCZ8 p-value	0.567	0.050
SCZ4 vs. SCZ5 p-value	0.043	0.277
SCZ4 vs. SCZ6 p-value	0.137	0.845
SCZ4 vs. SCZ7 p-value	0.675	0.012
SCZ4 vs. SCZ8 p-value	0.714	0.842
SCZ5 vs. SCZ6 p-value	0.535	0.179
SCZ5 vs. SCZ7 p-value	0.084	0.0004
SCZ5 vs. SCZ8 p-value	0.042	0.365
SCZ6 vs. SCZ7 p-value	0.268	0.016
SCZ6 vs. SCZ8 p-value	0.178	0.682
SCZ7 vs. SCZ8 p-value	0.895	0.006

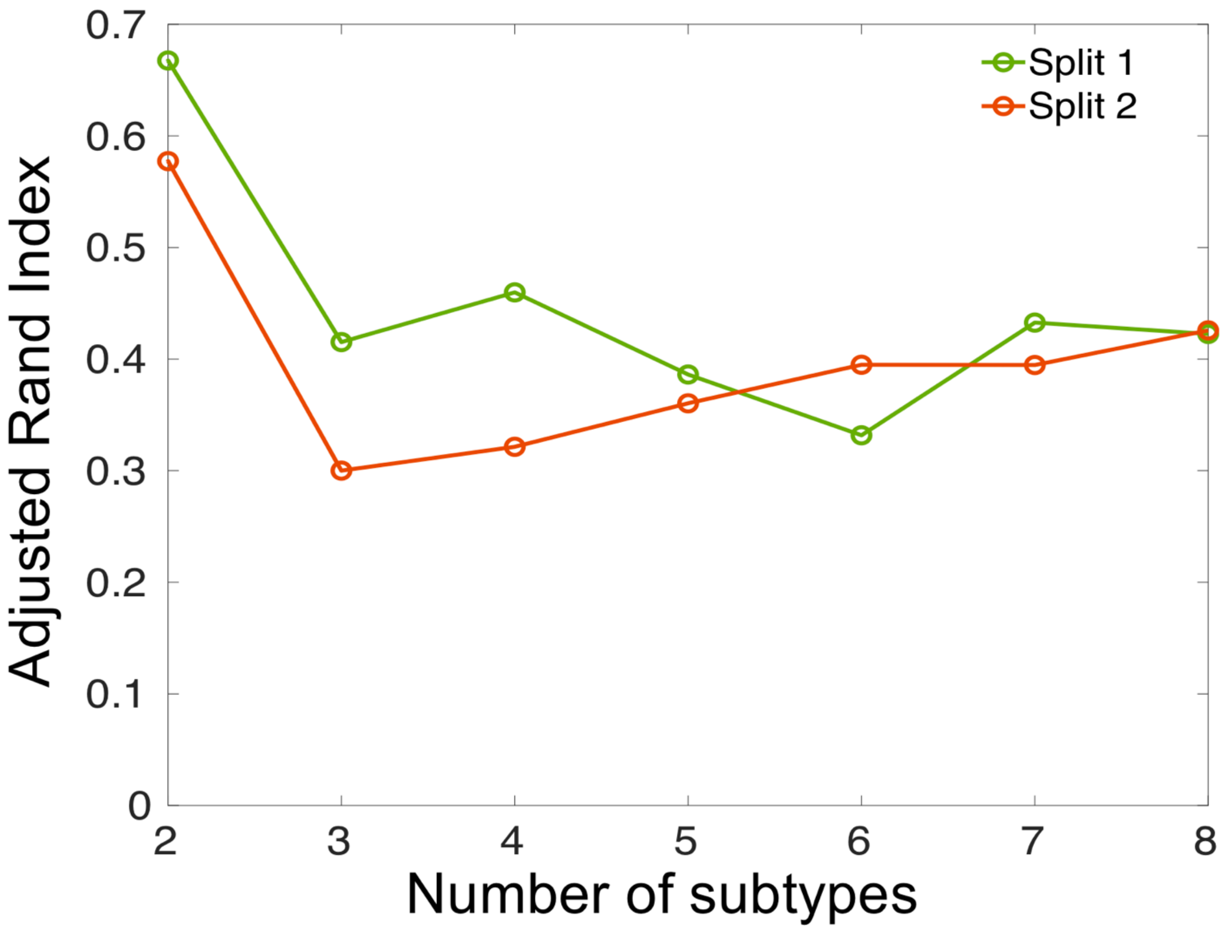

Figure S3: Cross-validated stability of split-half samples: Adjusted Rand Index (ARI) vs. number of subtypes (K) indicating that K = 2 yields highly reproducible subtypes in both Split 1 and Split 2.

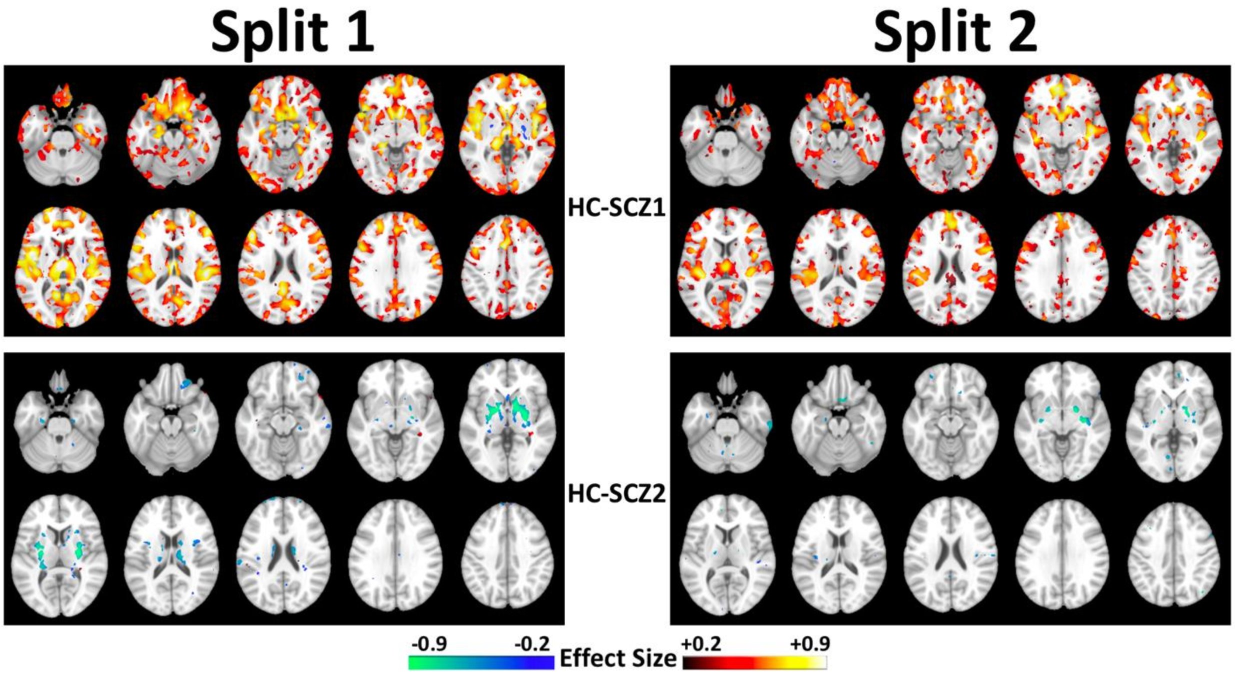

Figure S4: GM volumetric differences between each subtype and HC for K = 2 in Split 1 (left column) and Split 2 (right column). In both splits, the GM volumetric patterns (FDR-p < 0.05) are similar to the ones obtained using the full sample.

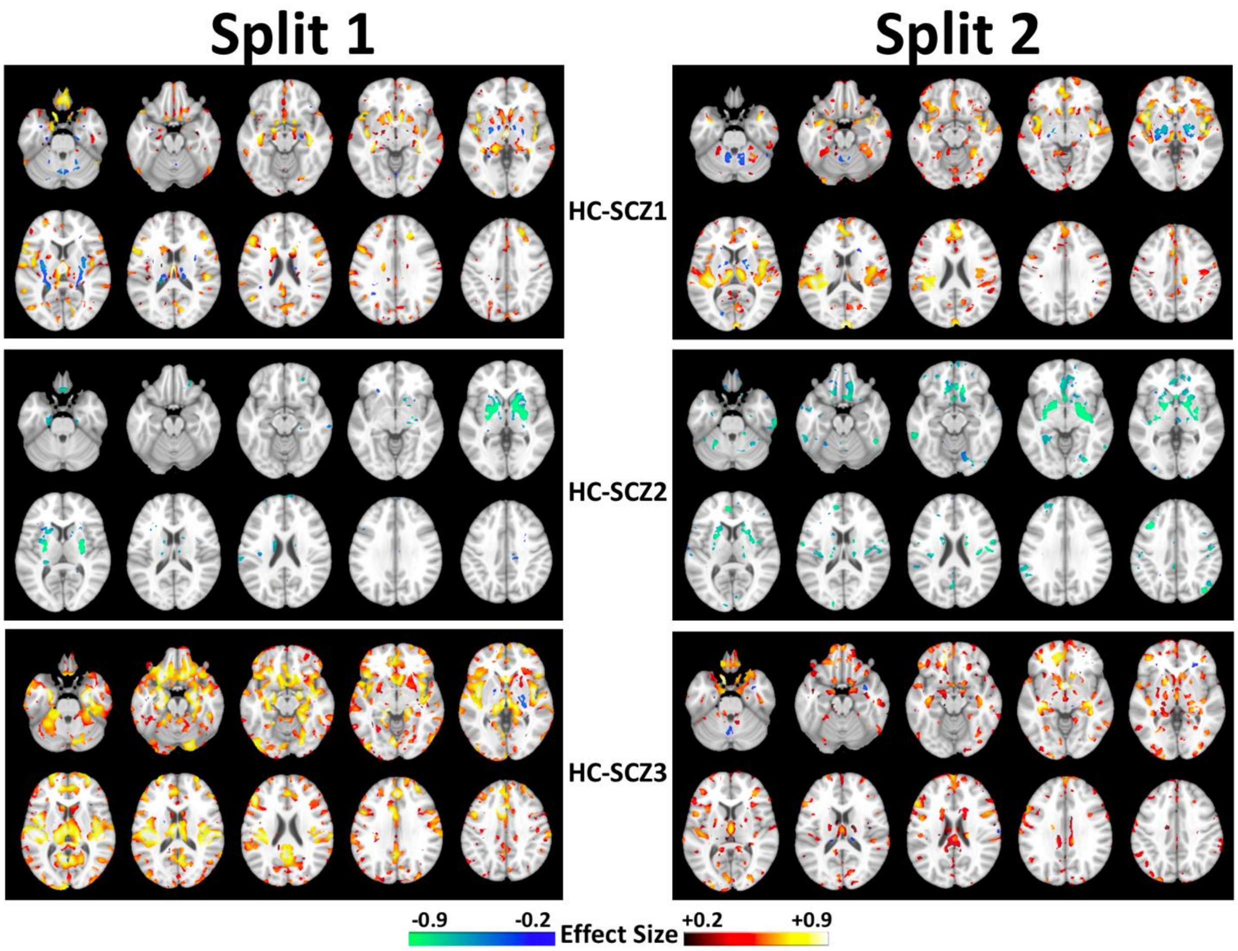

Figure S5: GM volumetric differences between each subtype and HC for K = 3 in Split 1 (left column) and Split 2 (right column). The volumetric profiles (FDR- $p < 0.05$) are not reproducible between Split 1 and Split 2.

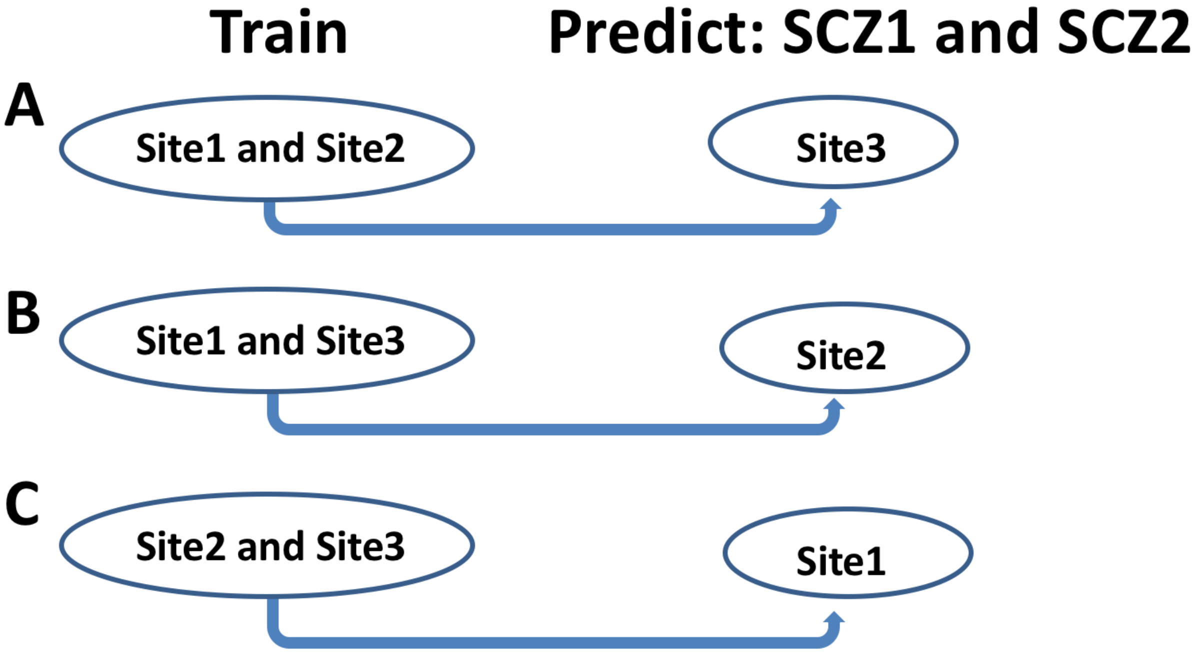

Figure S6: Schematic of the leave-one-site-out prediction: The two subtypes (SCZ1 and SCZ2) of each site were determined using the HYDRA-models trained on the other two sites. The percentage overlap of patients that were assigned to the same subtype was 86.72% (83.33% in Site1, 86.21% in Site2 and 90.63% in Site3) when compared with the original assignments obtained by taking all the sites together.

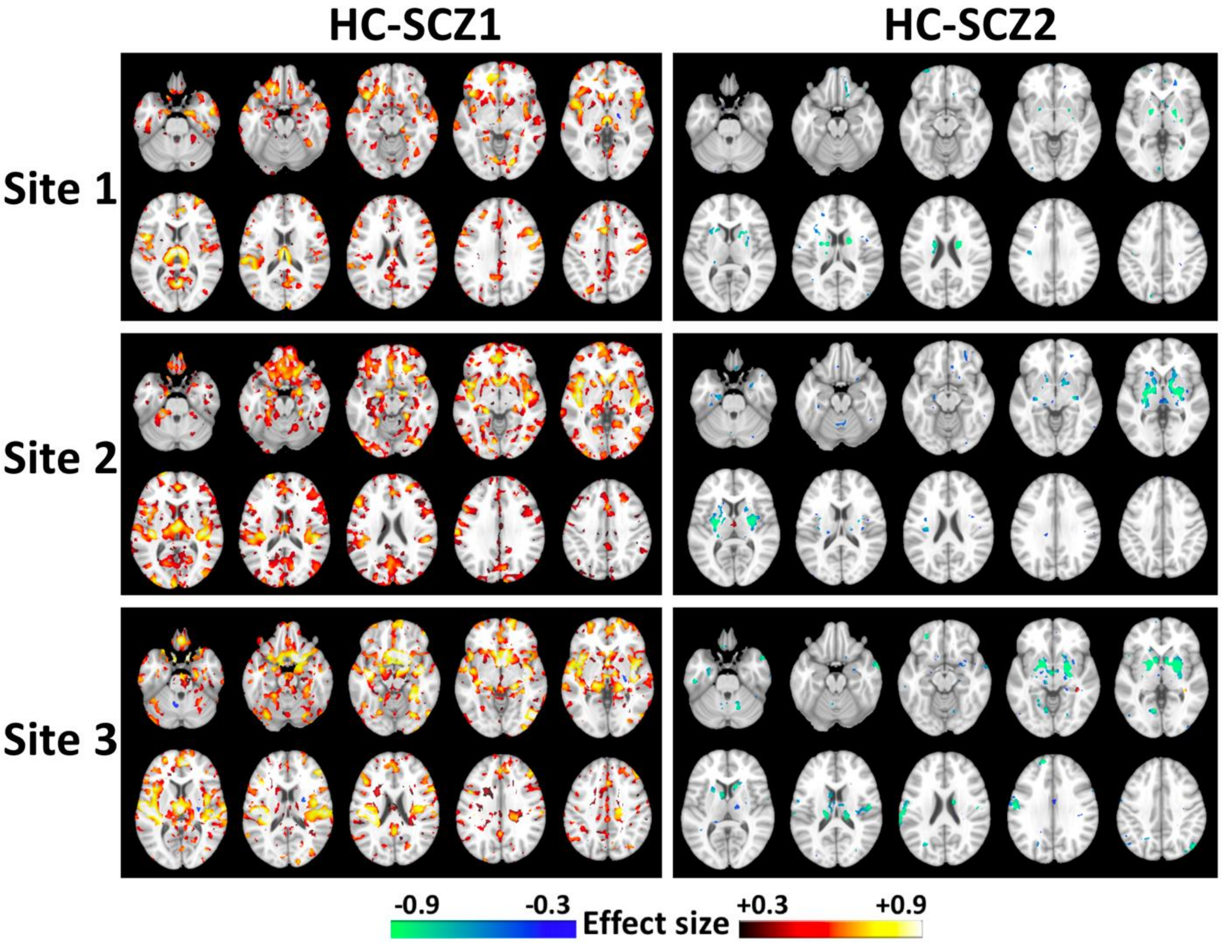

Figure S7: GM volumetric differences between each subtype and HC in each site**:** Compared to HC, SCZ1 shows widespread smaller volumes prominently in the thalamus, nucleus accumbens, medial temporal, medial prefrontal/frontal and insular cortices, and SCZ2 shows larger volume in the basal ganglia. The displayed results are (*FDR-p* < 0.05).

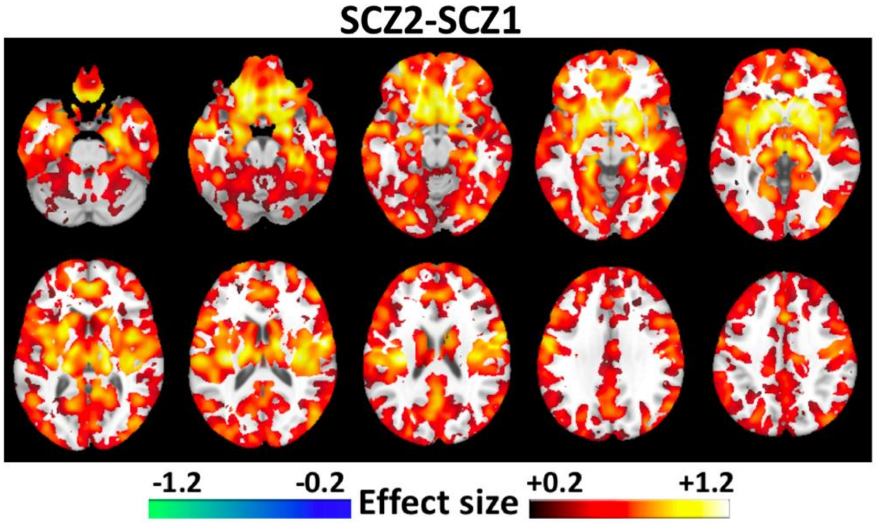

Figure S8: Comparison of GM volumetric patterns between the two subtypes (FDR-p < 0.05).

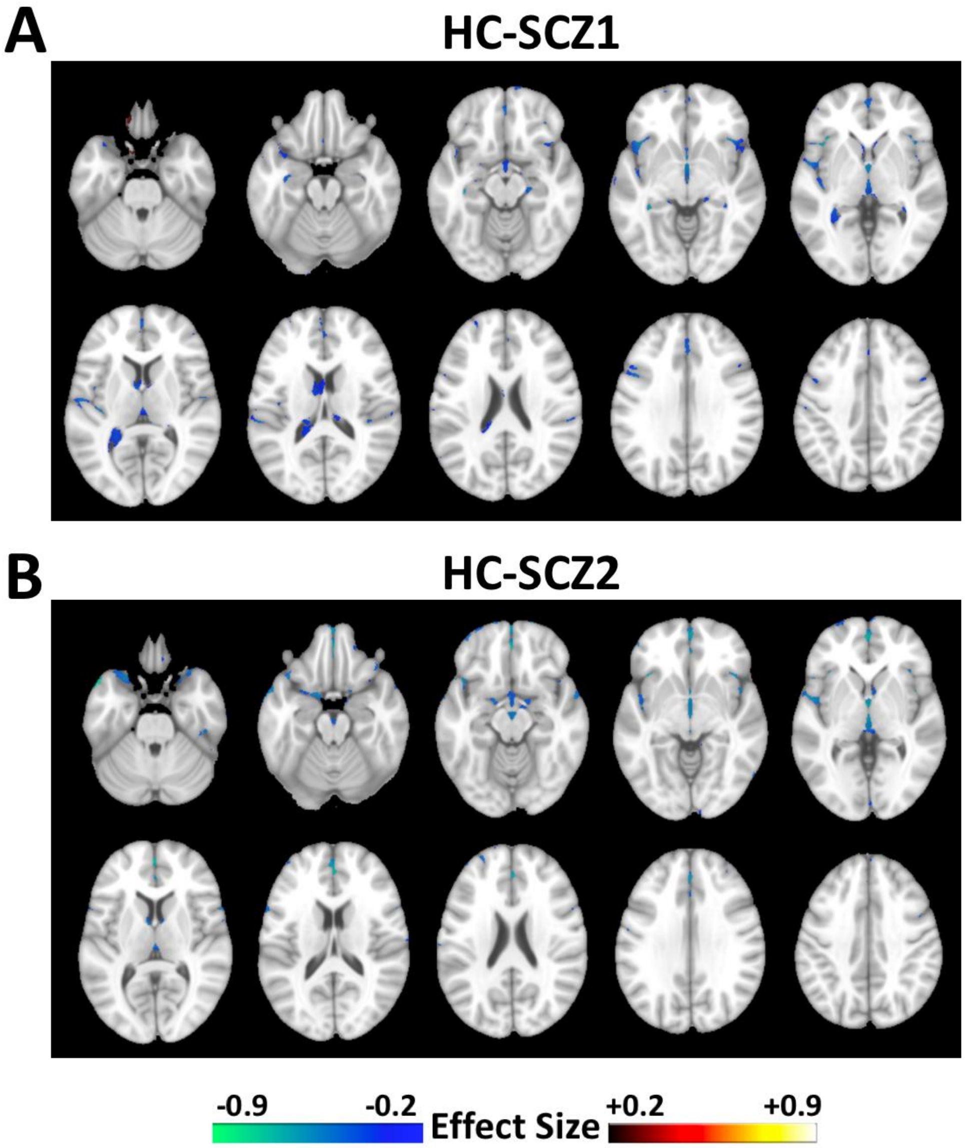

Figure S9: Patterns of CSF volumetric differences between each subtype and HC (FDR-p < 0.05).

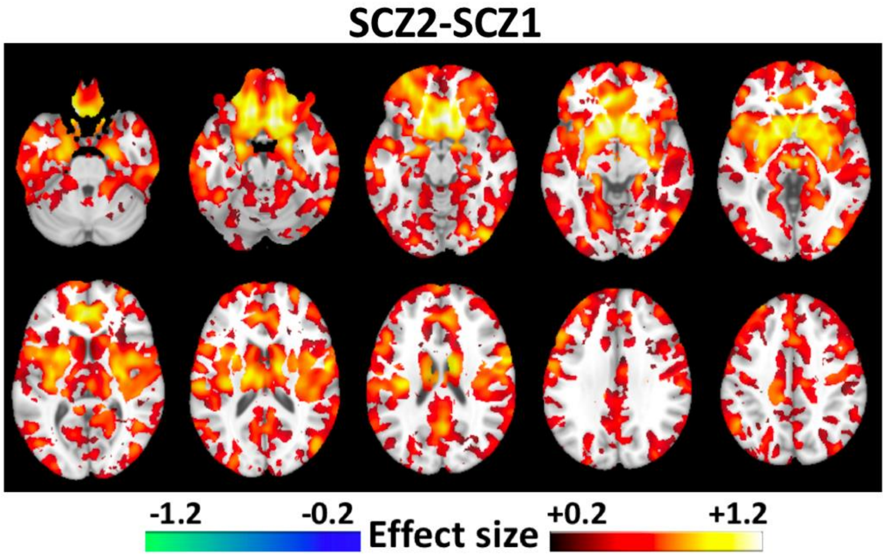

Figure S10: Comparison of GM volumetric patterns between the two subtypes estimated using leave-one-site-out (FDR-p < 0.05). These results are consistent with those obtained using the entire sample together (Figure S8).

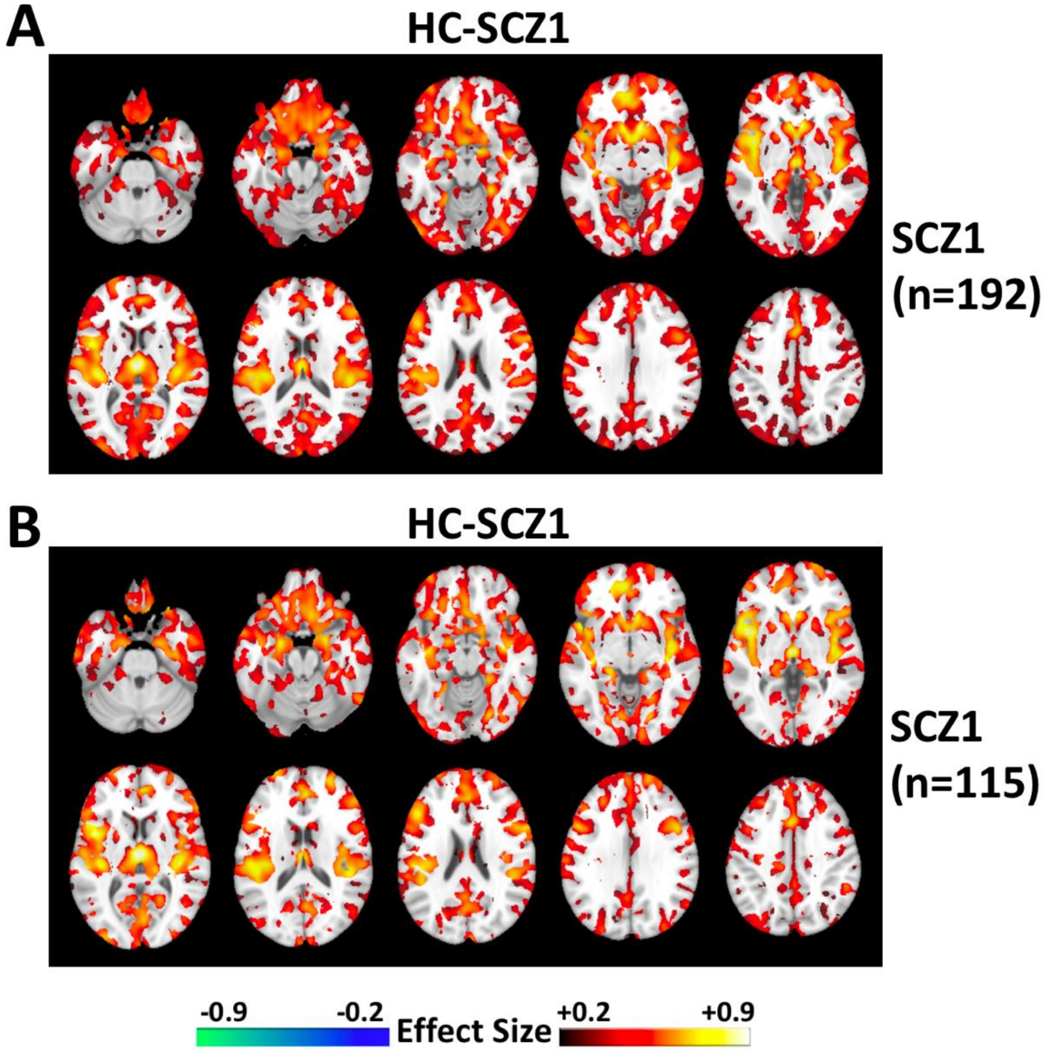

Figure S11: GM volumetric differences of SCZ1 compared with HC for **A**) full sample size and **B**) a subsample of the same size as that of SCZ2 (FDR-p < 0.05 results are displayed). This experiment indicates that the finding of smaller GM volumes observed in SCZ1 is not explained by the larger sample size of SCZ1.

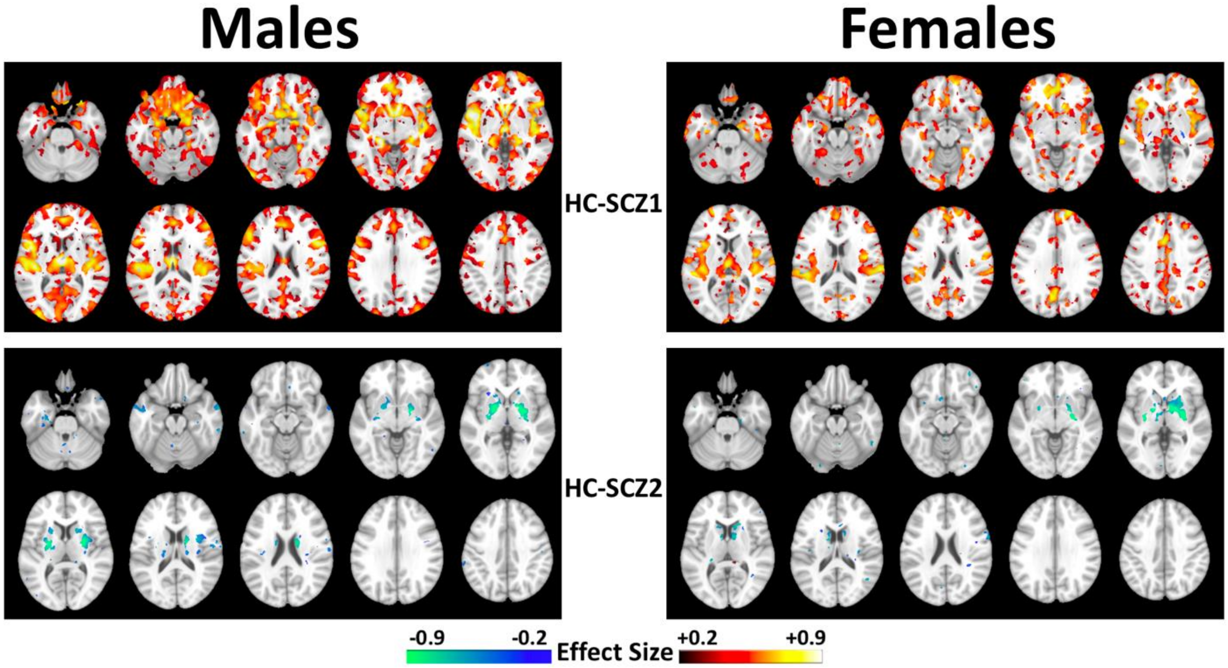

Figure S12: GM volumetric patterns of each subtype relative to HC in male (left column) and female (right column) subjects, separately (FDR- $p < 0.05$). The patterns are consistent with the overall mixed-sex pattern of volumetric differences, indicating that the subtype estimation was not driven by sex differences in the two subtypes.

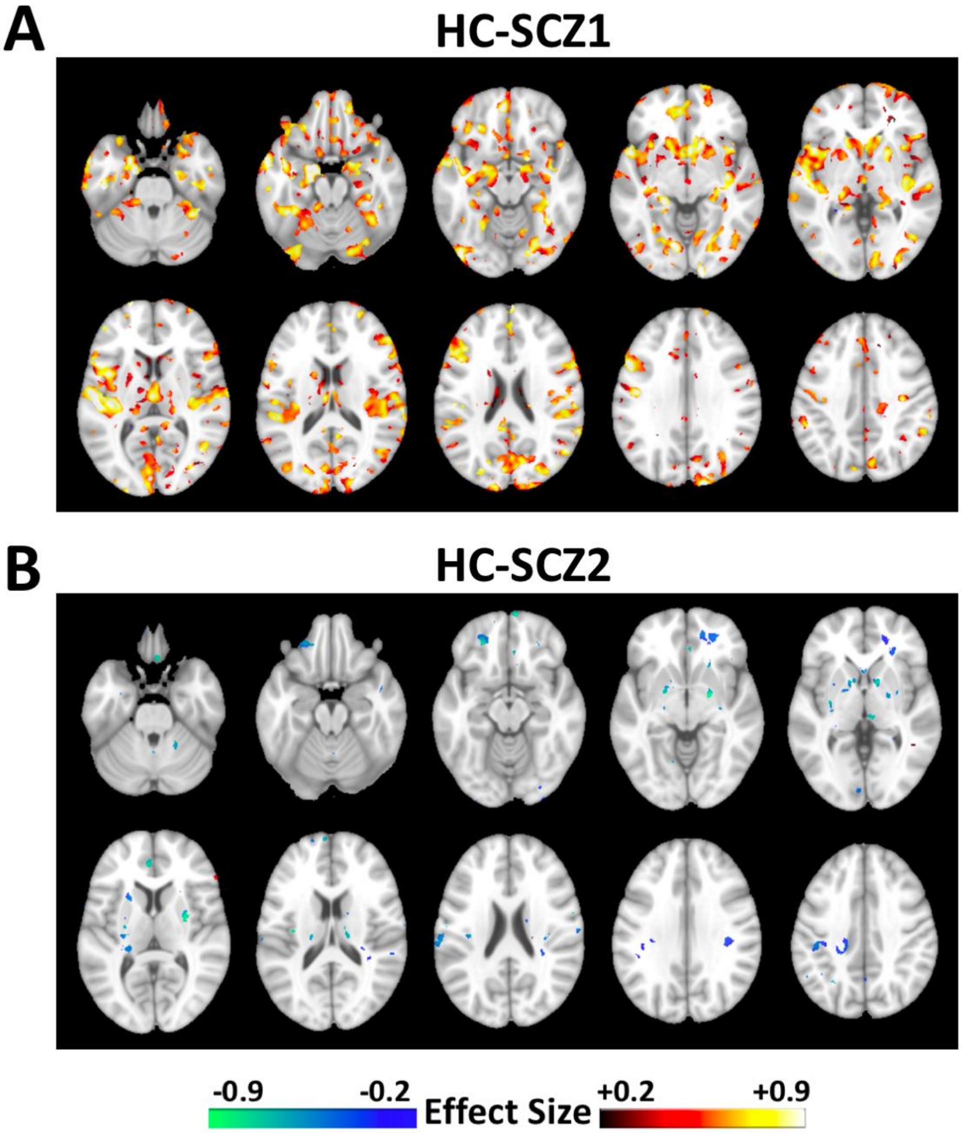

Figure S13: Altered GM volumetric patterns of K = 2 Subtypes compared with HC after adjusting for CPZ-equivalent dose, for a subset of patients with CPZ-equivalent dose data [n = 125 SCZ1 and n = 87 SCZ2]. These patterns are consistent with those obtained from the entire sample, albeit weaker, largely due to the smaller sample size (FDR- $p < 0.05$).

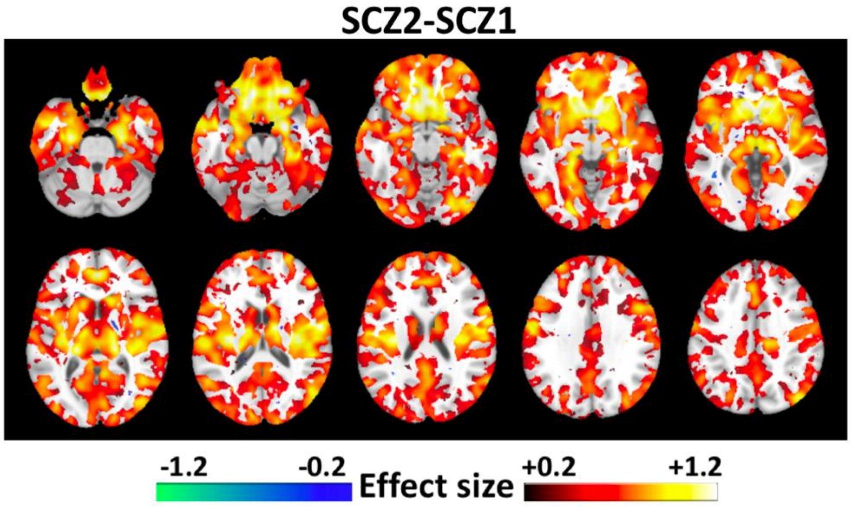

Figure S14: Comparison of GM volumetric differences between the two subtypes, after adjusting for CPZ-equivalent dose (FDR-p < 0.05). These results are consistent with those obtained without CPZ adjustment (Figure S8 and Figure S10).

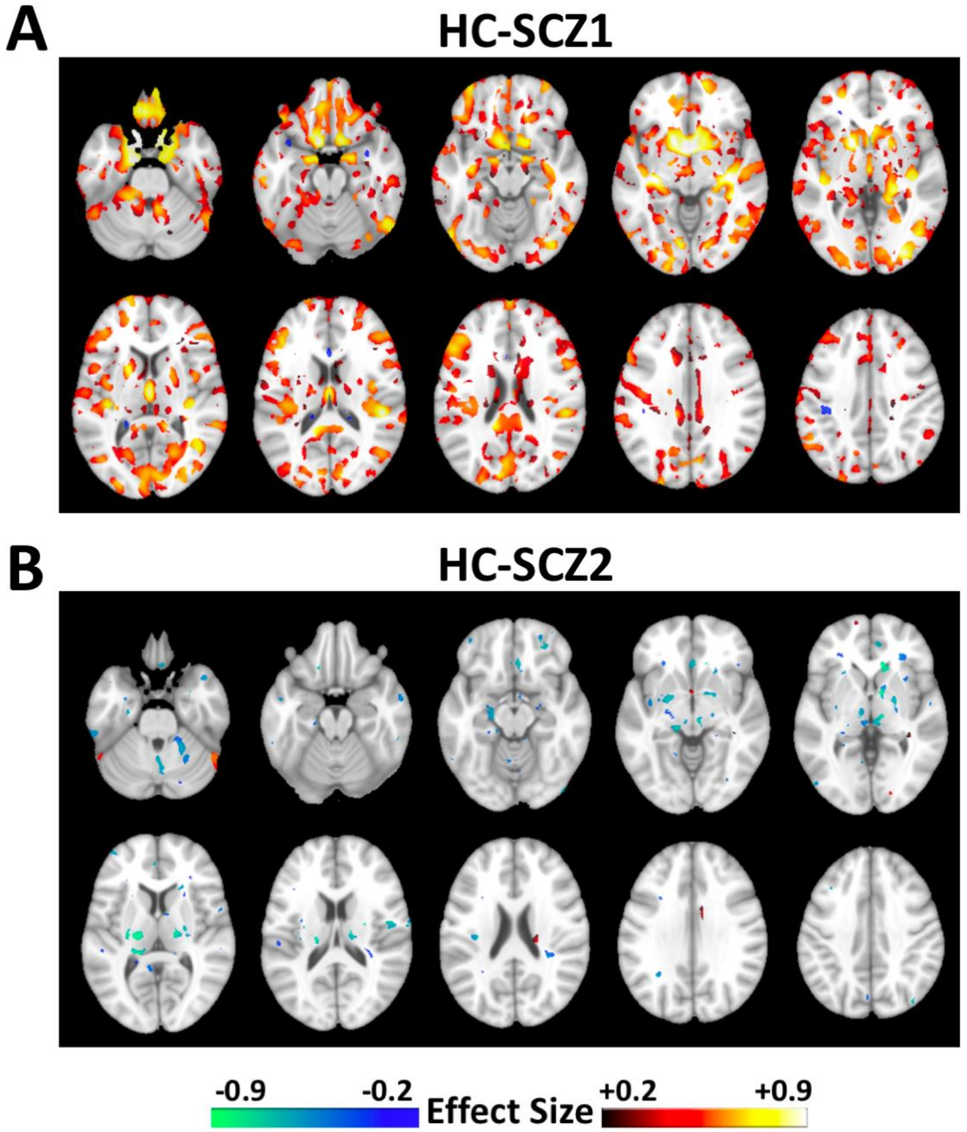

Figure S15: GM volumetric differences between each subtype and HC, restricted to patients who had illness duration less than 2 years (0.54 years average) (FDR-p < 0.05). The patterns are consistent with the findings from the larger group, except somewhat weaker, largely due to the smaller sample size.

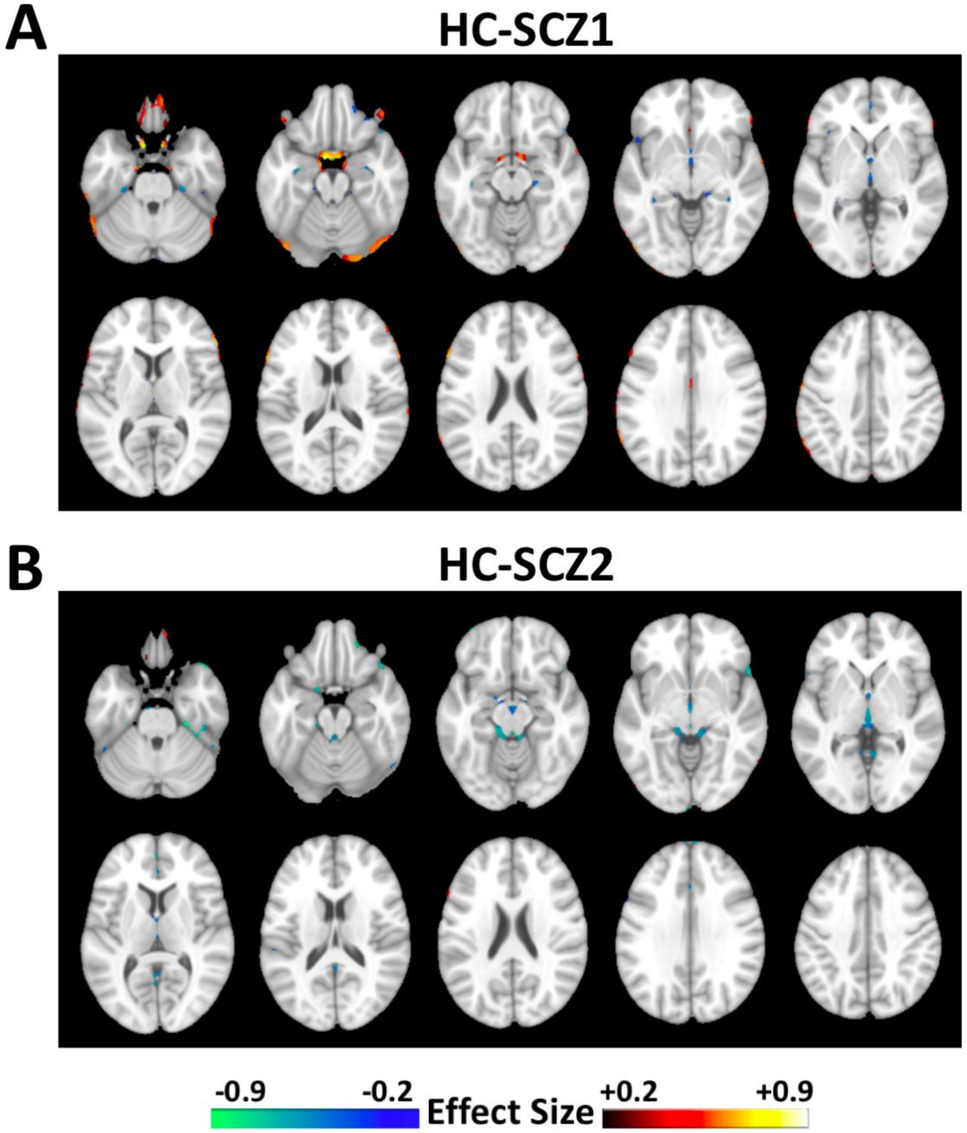

Figure S16: Patterns of CSF volumetric differences between each subtype and HC, restricted to patients who had illness duration less than 2 years (FDR-p < 0.05).

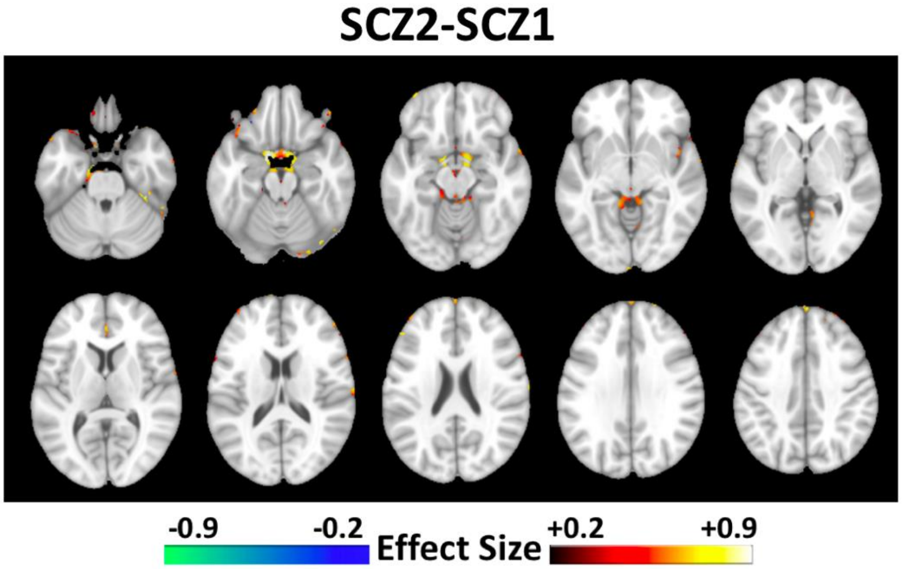

Figure S17: Comparison of CSF volumetric patterns between the two subtypes, restricted to patients who had illness duration less than 2 years (FDR-p < 0.05).

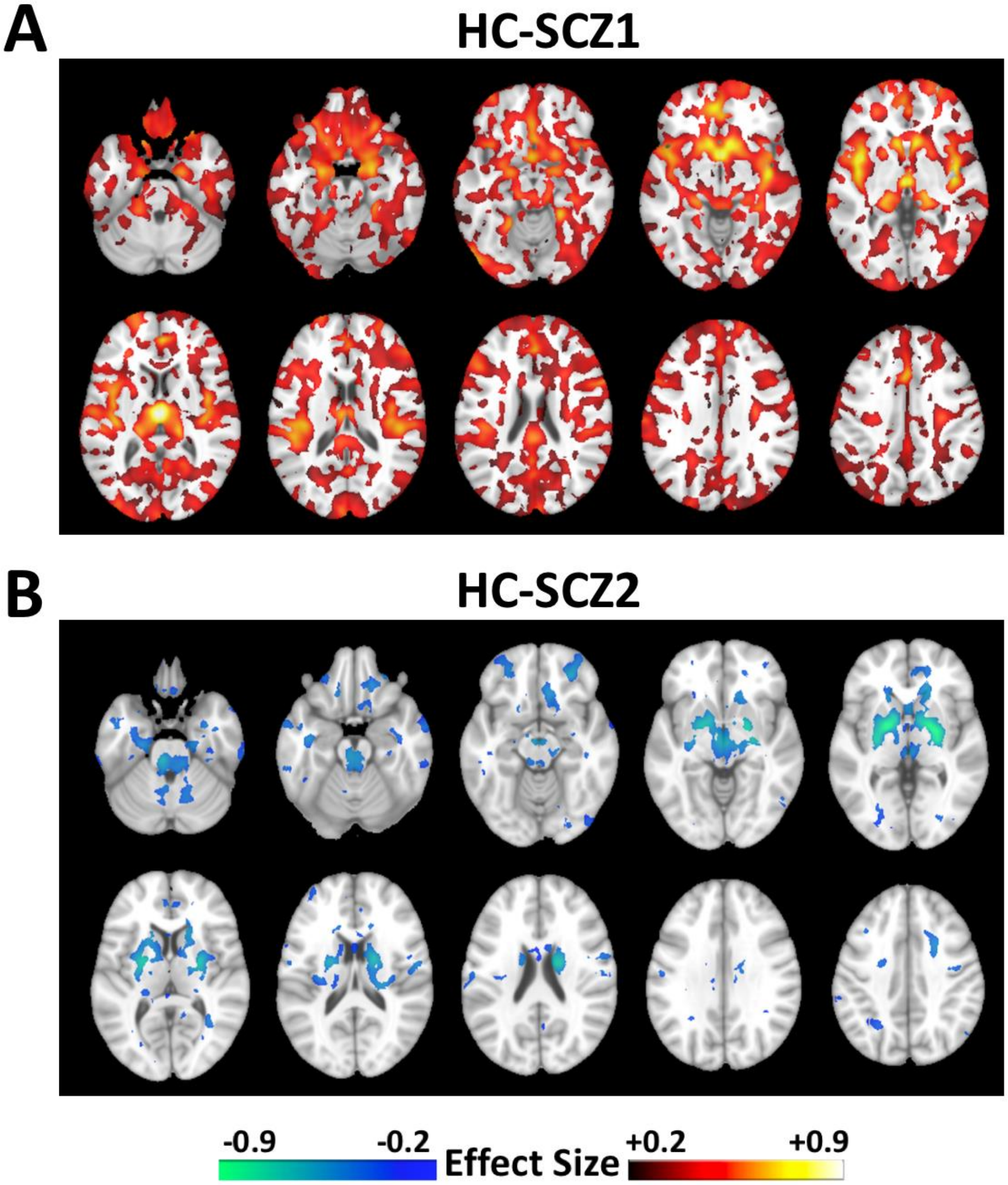

Figure S18: Patterns of total brain tissue (GM+WM) volumetric differences between each subtype and HC (FDR-p < 0.05). The results are consistent with GM and WM comparisons separately, thereby bolstering our confidence that the MRI contrast between GM and WM did not influence the tissue segmentation and then the clustering results.

References

Andreasen NC. The scale for the assessment of negative symptoms (SANS). University of Iowa, Iowa City 1983.

Andreasen NC. The scale for the assessment of positive symptoms (SAPS). University of Iowa, Iowa City 1984.

Andreasen NC. The scale for the assessment of negative symptoms (SANS): conceptual and theoretical foundations. Br J Psychiatry 1989; 155: 49-52.

Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics Surveys 2010; 4(0): 40-79.

Ben-Hur A, Elisseeff A, Guyon I. A stability based method for discovering structure in clustered data. Pac Symp Biocomput 2002: 6-17.

Davatzikos C, Genc A, Xu D, Resnick SM. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. NeuroImage 2001; 14(6): 1361-9.

Doshi J, Erus G, Ou Y, Resnick SM, Gur RC, Gur RE*, et al.* MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. NeuroImage 2016; 127: 186-95.

Hubert L, Arabie P. Comparing partitions. J Classif 1985; 2(1): 193-218.

Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987; 13: 261-76.

Kulesza A, Taskar B. Determinantal Point Processes. arXiv preprint arXiv:12076083 2012.

Lange T, Roth V, Braun ML, Buhmann JM. Stability-based validation of clustering solutions. Neural Comput 2004; 16: 1299-323.

Nichols TE, Homes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 2001; 15: 1-25.

Ou Y, Sotiras A, Paragios N, Davatzikos C. DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. Med Image Anal 2011; 15(4): 622-39.

Rozycki M, Satterthwaite TD, Koutsouleris N, Erus G, Doshi J, Wolf DH*, et al.* Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. Schizophr Bull 2018; 44(5): 1035-44.

Satterthwaite TD, Wolf DH, Loughead J, Ruparel K, Valdez JN, Siegel SJ*, et al.* Association of enhanced limbic response to threat with decreased cortical facial recognition memory response with schizophrenia. Am J Psychiatry 2010; 167: 418-26.

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA*, et al.* N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010; 29(6): 1310-20.

van Erp TG, Preda A, Nguyen D, Faziola L, Turner J, Bustillo J*, et al.* Converting positive and negative symptom scores between PANSS and SAPS/SANS. Schizophr Res 2014; 152(1): 289-94.

Varol E, Sotiras A, Davatzikos C. MIDAS: Regionally linear multivariate discriminative statistical mapping. NeuroImage 2018; 174: 111-26.

Varol E, Sotiras A, Davatzikos C, Alzheimer's Disease Neuroimaging I. HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. NeuroImage 2017; 145(Pt B): 346-64.

Wolf DH, Satterthwaite TD, Kantrowitz JJ, Katchmar N, Vandekar L, Elliott MA*, et al.* Amotivation in schizophrenia: integrated assessment with behavioral, clinical, and imaging measures. Schizophr Bull 2014; 40(6): 1328-37.

Zhang T, Koutsouleris N, Meisenzahl E, Davatzikos C. Heterogeneity of structural brain changes in subtypes of schizophrenia revealed using magnetic resonance imaging pattern analysis. Schizophr Bull 2015; 41(1): 74-84.

Zhu J, Zhuo C, Liu F, Xu L, Yu C. Neural substrates underlying delusions in schizophrenia. Sci Rep 2016; 6: 33857.

Zhuo C, Ma X, Qu H, Wang L, Jia F, Wang C. Schizophrenia patients demonstrate both inter-voxel level and intra-voxel level white matter alterations. PLoS One 2016; 11(9): e0162656.