Neurobiological Divergence of the Positive and Negative Schizophrenia Subtypes Identified Upon a New Factor-Structure of Psychopathology Using Non-Negative Factorization: An International Machine-Learning Study

Supplemental Information

Contents

Detailed sample information for the international dataset 4

Supplemental Methods and Results 7

Methodological details and model evaluation strategies for OPNMF 7

Model evaluation for selecting the optimal number of factors 9

Methodological notes for bootstrapping and 10-fold cross-validation based evaluations 17

Cross-sample analysis after accounting for sample size, age and illness duration differences 18

Evaluations on the pooled sample 19

Evaluation of loading and item-score predictions 19

Outlier Detection 20

Relationship among variables and factor analysis comparison 23

Inter-correlations of OPNMF factor-loadings 23

Additional ANOVA analyses 23

Inter-item correlations, and correlations between OPNMF factors and PANSS subscales 25

Exploratory and confirmatory factor analysis 26

Quantitative comparison of PCA and OPNMF factor models 27

Identification of psychopathological subtypes 28

Methodology and cluster selection 28

Assessment of clustering stability 30

Comparison of ambiguous class with core subtypes 31

Longitudinal stability analysis 32

Additional clustering analyses 34

Classification of psychopathological subtypes based on resting-state functional connectivity 37

MRI data acquisition and preprocessing 37

Connectivity matrix construction and classification analysis 39

Functional characterization analysis 42

Supplemental Tables 43

Table S1. 43

Table S2. 45

Table S3A. 46

Table S3B. 47

Table S4. 48

Supplemental Figures 51

Figure S1. 51

Figure S2 52

Figure S3 53

Figure S4. 54

Figure S5 55

Figure S6A-S6B. 56

Figure S7A-S7B. 58

Figure S8 60

Figure S9A-S9B. 61

Figure S10 63

Figure S11 64

Figure S12. 65

Figure S13. 65

Figure S14. 66

Figure S15 67

Figure S16 68

Figure S17 69

Figure S18. 70

Figure S19 71

Figure S20 72

Figure S21 73

Figure S22 74

Figure S23 75

Figure S24. 76

Figure S25 ..... **77**

Figure S26 ..... **78**

Supplemental References: **79**

Detailed sample information for the international dataset

The Utrecht sample:

Patients with chronic schizophrenia were diagnosed according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria (1) by an independent psychiatrist using the "Comprehensive Assessment of Symptoms and History (CASH)" (2). This study was approved by the Humans Ethics Committee of the University Medical Center Utrecht with written informed consent obtained from the all participants (3).

The Göttingen sample:

Patients were recruited from the Department of Psychiatry and Psychotherapy, University Medical Center Göttingen. They met the diagnostic criteria of schizophrenia according to DSM-IV (1). Patients who had substance abuse within the last month, cannabis abuse within the last 2 weeks, past or present substance dependency, somatic or mental disorders that would interfere with the protocol, acute suicidal tendency or an inability to give written consent were excluded (4).

The Groningen sample:

Diagnosis of schizophrenia was established based on the DSM-IV criteria (1), confirmed by a Schedules for Clinical Assessment in Neuropsychiatry (SCAN) interview (5). Exclusion criteria included a personal or family history of epileptic seizures, a history of significant head trauma or neurological disorder, the presence of intracerebral or pacemaker implants, inner ear prosthesis or other metal prosthetics/implants, severe behavioral disorders, current substance abuse, and pregnancy (6). The study was approved by the Institutional Review Board of the University Medical Center Groningen.

The Lille sample:

Patients were diagnosed with schizophrenia according to the DSM-IV-TR criteria (7). All patients routinely presented frequent (more than 10 per day) and resistant hallucinations as evaluated with item P3 of the PANSS. The exclusion criteria included the presence of an Axis-II diagnosis, secondary Axis-I diagnosis, neurological or sensory disorder, and a history of drug abuse, which was based on a clinical interview and urine tests that were administered at admission. The study was approved by the local ethics committee (CPP Nord-Ouest IV, France).Written informed consent from each patient was obtained (8).

The Munich sample:

All participants provided informed consent in accordance with the Human Research Committee guidelines of the Klinikum Rechts der Isar, Technische Universität München. Patients with a diagnosis of schizophrenia based on the Structured Clinical Interview for DSM-IV (SCID-I German version) were recruited from the Department of Psychiatry, Klinikum Rechts der Isar, TU München. Exclusion criteria were current or past neurological or internal systemic disorder, current depressive or manic episode, substance misuse (except for nicotine) and cerebral pathology on MRI (9,10).

The Albuquerque sample：

This dataset was collected and shared by the Mind Research Network and the University of New Mexico funded by a National Institute of Health Center of Biomedical Research Excellence (COBRE; http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html). Patients with schizophrenia were diagnosed based on DSM-IV using the Structured Clinical Interview used for DSM-IV axis I disorders (SCID). Informed consent was obtained from participants at the University of New Mexico. All patients were chronic and with relatively well-treated symptoms by a variety of antipsychotic medications (no medication changes in 1 month). Those patients with a history of neurological disorder, head trauma with loss of consciousness greater than 5 min, mental retardation, active substance dependence or abuse (except for nicotine) within the past year, current use of mood stabilizers, history of dependence on PCP, amphetamines or cocaine, or history of PCP, amphetamine, or cocaine use within the last 12 months were excluded (11).

The Wayne State sample:

Diagnosis of schizophrenia was established according to the DSM-V criteria (12) using the Structured Clinical Interview used for DSM-V axis I disorders. The Wayne State University Institutional Review Board approved all experimental procedures, and written informed consent was obtained from each patient. All patients were on stable antipsychotic treatments with either first generation, second generation antipsychotics or a combination of both. Exclusion criterion: (i) significant history of, or current medical or neurologic illness requiring systemic treatment; (ii) neurologic disorders, including head injury with loss of consciousness; (iii) Significant drug or alcohol use in the previous month or meeting DSM-V criteria for substance Abuse; (iv) meeting the DSM-V criteria for schizoaffective disorder or any other psychotic disorders other than schizophrenia; (v) co-morbidity for any (major) DSM-V Axis I diagnosis.

The Aachen sample:

Patients were diagnosed with schizophrenia according to the DSM-IV criteria (1) using the German version of the Structured Clinical Interview for DSM Disorders (SCID) by attending psychiatrists. Any patients with past or current presence of secondary Axis-I diagnosis, neurological or sensory disorder, and a history of drug abuse were excluded. The study was approved by the ethics committee of the Medical Faculty of the RWTH Aachen University with written informed consent obtained (13,14).

The Singapore sample:

Patients were diagnosed with schizophrenia according to the DSM-IV criteria (1) using the Structured Clinical Interview for DSM-IV Axis I disorders-Patient Edition (SCID-P) (15) by the treating psychiatrist. Participants were free from a history of neurological illness or a diagnosis of alcohol or drug misuse in the past three months based on DSM-IV criteria (1). All patients were on a stable dose of antipsychotic medication for at least 2 weeks, and none had medication withdrawn for the purpose of the study (16).

Detailed clinical characteristics for the international sample can be found in Table S1.

Supplemental Methods and Results

Methodological details and model evaluation strategies for OPNMF

Non-negative matrix factorization (NMF) produces a factorization that both of the factors (basis vectors) and the factor-loadings contain no negative elements. This parts-based learning approach models data by additive combinations of non-negative basis vectors, making the factorization results can be intuitively interpreted. NMF and its variants have been widely used in recent biomedical studies, including metagene discovery, functional characterization of genes, identification of structural brain networks, and cancer subtypes stratification (17-20). The present study is the first practice of applying this promising method to the PANSS data in psychosis to discover the latent dimensional structure of psychopathology. NMF is typically achieved by solving the following energy minimization problem:

$$\min \ \|V - WH\|_F$$
$$s.t. W \geq 0;$$

where *W* is the basis matrix (*m* attributes × *r* latent factors) containing the parts information. *H* is the (*r* factors x *n* data instances) matrix containing the loading coefficients, when used together with *W*, approximate the data matrix *V*.

Upon the above initial NMF algorithm, many studies have been devoted to developing diversiform extensions of NMF with different constraints to achieve specific practical purposes. One of the principal aspects and what we are most interested in factorizing the PANSS data is to expect a sparse representation of schizophrenia psychopathology. Sparsity representation could provide almost clustering like structure that facilitates the determination of item-assignment to specific dimensions but also retain the weights information of an item in belonging to each of the dimensions. Moreover, we would expect that the current defined factor-structure can be generalized to novel samples. In these considerations, we adopted a variant of NMF, namely the orthonormal projective NMF (OPNMF) (20), to discover the latent structure of the PANSS. This method differs from the original NMF in that it replaces the loading matrix by the inner product of the basis vectors and the input data matrix, making OPNMF to be projectable [i.e., $H = W^T V$, and thus the dictionary *W* (basis matrix) can be readily generalized to new data]. Due to the projective constraint in OPNMF, the loading coefficients are not free variables any

more, which facilitates each factor to focus on specific parts of the data, leading to factors that are sparse, overlap less and are naturally more orthogonal. This is important as we could extract more compact and homogeneous latent factors from the PANSS data. The additional orthonormality constraint promotes the orthogonality between the learned factors. Sparsity is of great importance in signal decomposition and biological interpretation (21) and has been associated with improved generalizability (22). In contrast to other NMF variants to achieve sparsity, OPNMF does not involve any regularization terms or trade-off parameters, but is still able to learn more spatially localized, parts-based representations of the imported data patterns. Importantly, by enforcing the orthonormality constraint, the multiplicative update step becomes simpler which leads to less computational expense. This allows us to converge better to a local minimum and facilitates the implementation of various cross-validation and out-of-sample generalization evaluations to get a stable and robust pattern of the latent factor structure of the PANSS.

The whole optimization process for OPNMF is to minimize the reconstruction error measured by frobenius norm between the input data matrix *V* and its estimate by only updating the basis matrix *W*:

$$\min \ \|V - WW^T V\|_F$$
$$s.t. W \geq 0; WW^T = \boldsymbol{I}$$

where matrix *W* conveys factor information in each column with respect to the co-occurrence properties of the PANSS items which has a size of *m (item)* x *r* (*r* is the number of the estimated factors with each of the *r* columns defining a psychopathological dimension). Entry w_{ij} of *W* is the coefficient of item *i* in factor *j*. WW^T is the projection matrix, on which the matrix *V* can project to yield a subspace so that to approximate itself. In this form, the loading matrix *H* is replaced by W^TV so that the basis matrix *W* can be used to represent new data. The orthonormality constraint of $WW^T=I$ requires *W* to be an orthonormal matrix, and the orthogonality between the vectors in the learned *W* yields sparse factors, i.e., dimensions of psychopathology. *H* then encodes the symptomatology of a given patient along the dimensions spanned by the basis matrix *W*, which has a size of *r* x *n* (each of the *n* columns represents the expressed symptomatic severity for a patient corresponding to the *r* factors defined in *W*) with entry h_{jk} represents the expression level of factor *j* in patient *k* that can be used for patient-centric analyses, e.g., clustering patients into subtypes.

The non-convex problem was approached by iteratively performing the below multiplicative update rule:

$$W_{ij} = W_{ij} \frac{(VV^T W)_{ij}}{(WW^T VV^T W)_{ij}}$$

This update rule guarantees the positivity of the estimated factors, while monotonically decreasing the energy towards attaining a local optimum. Proofs of convergence have been presented in detail in previous literature (23,24).

Choose of initialization method is important since a suitable initialization of *W* will facilitate fast convergence. Non-negative singular value decomposition (NNSVD) was employed here as the initialization strategy (25), which has the superiorities of reduced residual error, faster convergence than using random initialization (25), and most critically, renders the final non-negative decomposition to be deterministic.

Model evaluation for selecting the optimal number of factors

To determine the most robust, stable and generalizable factor model as a dimensionality reduced conceptualization of schizophrenia psychopathology, a sophisticated evaluation scheme was developed. Specifically, we employed three different data perturbation manners of split-half, bootstrap and 10-fold cross-validation to assess the resulting factor-models in two aspects (i.e., stability and generalizability). Of note, the below demonstrations are based on split-half cross-validation on each of the PHAMOUS and international samples independently, while the bootstrap and 10-fold cross-validation procedures on the respective samples, as well as the between-sample bootstrap-based comparison (PHAMOUS vs. international) analysis, are noted briefly in the "*Methodological notes for bootstrapping and 10-fold cross-validation based evaluations*" section with an emphasis on the points that are different from the split-half strategy.

In summary, the set of columns of the original item by patient matrix is randomly split into two independent sets of equal length. Then, OPNMF was performed on the ensuing two-half submatrices and each item was assigned to a certain factor for the submatrices. Three evaluation indices were calculated based on the similarities of item-assignment (adjusted Rand index [aRI], variation of information [VI], and Jaccard index [JI]), and one index (i.e., the concordance index [CI]) was based on

the whole entries of the basis matrix *W* (here instead of using the hard-assignment results of the items based on the entries of *W*, CI was calculated based on the initial values within *W*) between the two submatrices to demonstrate stability. Generalizability was evaluated by measuring increase in out-of-sample reconstruction error.

Overview

1. Stability:

A. based on hard-assignment of the items to specific factors (as a natural clustering).Three evaluation indices of aRI, JI and VI were employed to reflect how similar of the factor-label assignment for each item and item-pair grouping between the two split samples in each split-half realization.

B. based on the initial values of the all entries in the basis matrix *W*. Since an item can be influenced by multiple dimensions and may have small contributions to other factors (low coefficients loaded on other factors besides the one an item is assigned to), the CI which reflects the concordance of the cosine similarity for each pair of the PANSS items between the factorizations of split-samples was thus employed to account for the items with multiple factor-memberships.

2. Generalizability (indicates how well novel data can be compressed by a given dictionary):

Generalizability is assessed by measuring increase in out-of-sample reconstruction error. The reconstruction error is the absolute differences between the reconstructed matrix and the original data matrix. The out-of-sample increased reconstruction error refers to how much worse the matrix is reconstructed relative to the original data matrix by the dictionary (basis matrix) obtained from model-unseen sample comparing to the reconstruction error calculated by the matrix recovered from the within-sample dictionary.

Detailed implementation of the above evaluation processes for split-half cross-validation is presented as follows：

Stability evaluation

1.A Stability based on hard-assignment of the PANSS items

Decomposition of the data matrix *V* results in two matrices, i.e. with the loading matrix W^TV we can cluster on it using cardinal clustering methods to identify the subtypes of schizophrenia patients according to their differential symptomatic expressions; the basis matrix *W* is exactly the factorization

results encoding the latent symptomatic dimensions of the PANSS. In this section, we focused on the basis matrix *W* to evaluate how many factors give the best low-dimensional presentation of the 30-item PANSS in schizophrenia. As that conducted in previous literature (26), we categorized the items into *k* factors based on the largest coefficients. Specifically, item *j* is placed in factor *i* if the w_{ij} is the largest entry in column *i*.

After assigning each item to a specific factor, we can choose the optimal K by using some well-established evaluation indices. First, the Jaccard index (JI) (27) was employed, which reflects the similarity between the factor-label assignment results from the factorizations of the two split-samples. The value of JI is computed as follows:

$$JI = \frac{|F1 \cap F2|}{|F1 \cup F2|} = \frac{|F1 \cap F2|}{|F1| + |F2| - |F1 \cap F2|}$$

where *F1* and *F2* represent the item-assignment results from factorizations of the respective two split submatrices. The JI value is derived as the ratio of the number of factor-assignment for the items common to both submatrices, divided by the total number of items present in both submatrices minus the numerator term. In this, JI value is bounded between 0 and 1, where a value of 1 indicates perfect correspondence between the item-assignment results of the two split halves, while 0 refers to no similarity.

A second metric of aRI (28) was further employed to complement JI by assessing the precision of data point assignment to the correct community, which better addresses specificity. ARI is a modified version of RI, which is adjusted for the chance placement of elements (i.e., penalizes for the placement of two data points from different true communities into the same community) and thus is stricter than RI with improved discrimination. The aRI can yield a value between -1 and +1, which has expectation 0 under the null hypothesis of randomness (the point assignment performance equivalent to random placement), and the negative values will happen if the index is less than the expected index. In our case, aRI was used as a measure of correspondence between the factorizations derived from the two split-samples, which is based on PANSS item-pair assignment to the factors. Higher values of aRI indicate better correspondence of item-pair placement between the two factor-models derived in one split-half realization, and a value of 1 represents identical item assignment. The equation is given as follows:

$$aRI = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$$

Assuming that we have two factorizations of *F1* and *F2*, and let denote the labels of the items for the two factorizations with *F1* = { *f1* } and *F2* = { *f2* }

$$aRI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where n_{ij}, a_i, b_j are values from the contingency table, n_{ij} denotes the number of item-pairs placed in common factors between factorizations *F1* and *F2* of which the items belong to the *i*-th factor of *F1* and to the *j*-th factor of *F2*

$$n_{ij} = |f1 \cap f2|$$

$$a_i = \sum_j n_{ij}$$

$$b_j = \sum_i n_{ij}$$

a_i refers to the number of items in *i*-th factor of *F1*, b_j refers to the number of items in *j*-th factor of *F2,* n = a_i= b_j. We computed the mean and median aRI across all split-half replications for each factor number *k*.

Besides the aforementioned two indices, we finally employed an information theory based criterion named variation of information (VI) (29) to estimate the stability of the factor solutions. This index has been widely used in the determination of optimal cluster number for cardinal clustering analyses in many biomedical studies (30). VI reflects the dissimilarity of item-assignment by quantifying the differential information between the two factor-solutions of split samples. By given the two split submatrices in one split-half realization, we compared the item-assignment results for the factorization

of one half sample (*F1*) to that derived from the other half (*F2*) with each *k* using the *VI* metric defined as follows:

$$VI_k(F1, F2) = H_k(F1) + H_k(F2) - 2I_k(F1, F2)$$

$H_k(F1)$ and $H_k(F2)$ are the entropies (the amount of information, here refers to the complexity of the hard-assignment results, i.e. larger factor number relates to higher complexity) of the hard-assignment results for the two split samples and I_k (*F1*, *F2*) is their mutual information (i.e., how much information one hard-assignment solution gives about the other). H_k and I_k can be computed as follows:

$$I_k(F1, F2) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} P(k_1, k_2) \cdot log \frac{P(k_1, k_2)}{P(k_1)P(k_2)}$$

$$H_k(F) = -\sum_{k=1}^{K} P(k) \cdot log\big(P(k)\big)$$

P(k) is the probability that an item belongs to factor *k* (column in the basis matrix *W*) and *P(k1,k2)* is the probability that an item belongs to factor k_1 in factorization (hard-assignment) result *F1* and factor k_2 in *F2*. *P(k)* and *P(k1,k2)* are computed according to:

$$P(k) = \frac{n_k}{n}$$

$$P(k_1, k_2) = \frac{\left|F1_{k_1} \cap F2_{k_2}\right|}{n}$$

Where n_k is the number of items in factor *k*, and n is the total number of items (i.e., 30). We computed the mean and median VIs across all split halves over the 10,000 replications for a given *k*. Lower VI refers to that the factor solution of a given factor number *k* estimated from the two split-samples share more information and thus in higher stability.

1.B Stability based on the whole entries of the basis matrix accounting for those items with multi-factor memberships

Since OPNMF not only generates almost clustering-like structure, but still allows small contributions from multiple items to a certain psychopathological dimension, we further evaluated the stability of the whole entries of the basis matrix *W* besides measuring the stability of hard-assigned items and item-pairs between the two split samples as aforementioned. The major motivation is that symptoms (as scored in each item) can overlap, i.e. an item can contribute to multiple psychopathological dimensions. Here, we introduced a method for evaluating the stability of the whole weights within *W* based on an item similarity matrix. First, we normalized W to $\overline{W}$ by dividing each row by its Euclidean norm. Then, we constructed the similarity matrix *S* by $\overline{W}^T\overline{W}$ which is a symmetric matrix with ones down the diagonal, and each entry represents the cosine similarity of two items given by the OPNMF decomposition. Equation for similarity matrix *S* calculation is given as below (see also Raguideau *et al.*[31])*:*

$$S_{jj'}^{W} = \frac{\sum_{l=1}^{m} w_{lj} w_{lj'}}{(\sum_{l=1}^{m} w_{lj}^2)^{1/2} (\sum_{l=1}^{m} w_{lj'}^2)^{1/2}}$$

Let *W* be a basis matrix with a dimension of *m* item x *r* factor, the similarity matrix S^W of *W* is a squared matrix with *r* columns. The *(j, j')* entry in S^W is the cosine of the angle between the column vectors *j* and *j'* of *W*.

Then, the concordance index (CI) is derived as *CI* = 1 - *D*, where *D* is the root mean squared difference (RMSE) between off-diagonal entries of the *S* matrices computed from the two split-samples (S_1 for W_1 which inferred from one half split-sample and S_2 for W_2 which inferred from the other half) in each split-half realization:

$$RMSE = \frac{1}{\sqrt{m(m-1)}} \|S_1 - S_2\|_F$$

Generalizability evaluation by assessing the increase in out-of-sample reconstruction error

Since the overall optimization goal for OPNMF is to minimize the cost function (namely the

reconstruction error), i.e., the squared Euclidean distance between the original data matrix and it's estimate, the measure of reconstruction error can be used to reflect how good a data matrix is reconstructed by a given basis matrix *W* with smaller reconstruction error corresponding to better low-rank approximation. In the light of this property, we used the metric of reconstruction error to evaluate how good a basis matrix captures the latent structure of an input PANSS matrix in the perspective that the better an independent sample can be compressed by a dictionary, the lower the reconstruction error will be. Specifically, we calculated the reconstruction error between the input data matrix and its low-rank approximation as their absolute arithmetic difference. Then, the absolute differences were summed up over items (for per subject) and were averaged over the subjects to derive the final metric for one split-half realization.

Furthermore, in the form of OPNMF, the basis matrix is projectable and thus can be generalized to novel data, i.e. a new matrix *X* can be reconstructed by using a given dictionary *W* as $X'=WW^TX$ with the corresponding reconstruction error $||X-X'||_F$. Likewise, smaller reconstruction error refers to better low-rank approximation for this basis matrix in the representation of new data. In the present study, we defined the transfer (out-of-sample increased) reconstruction error ($|V_1-W_2W_2^TV_1|-|V_1-W_1W_1^TV_1|$; V_1 and V_2 represent the two split-half submatrices) as a function of generalizability. This refers to how much worse the matrix is reconstructed relative to one half of the data (V_1) by the dictionary (W_2) obtained from the other half (i.e., $V_1-W_2W_2^TV_1$) comparing to the reconstruction error (i.e., $V_1-W_1W_1^TV_1$) calculated by the matrix recovered from the within-sample dictionary (W_1). These processes were repeated for each factor number *k* in the 10,000 split-half realizations. The most generalizable factor structure with rank *k* was selected when the median transfer reconstruction error of the all 10,000 split-half realizations is minimized, since a well-generalized factor-model should minimize the transfer reconstruction error in majority of the overall split-half realizations.

Detailed implementation of the current generalizability evaluation for split-half comparison is presented as follows:

First, the set of columns of the original data matrix is randomly split into two independent sets of equal length (split samples), and OPNMF was performed on the two split submatrices (V_1 and V_2) to derive the respective basis matrix (W_1 and W_2). Then, the two submatrices were reconstructed:

Reconstruct the submatrix V_1 by W_1:

$$\widehat{V_1}=W_1W_1^TV_1$$

Reconstruct the submatrix V_2 by W_2:

$$\widehat{V_2}=W_2W_2^TV_2$$

The within-sample reconstruction error for V_1 is:

$$RE_1=|\widehat{V_1}-V_1|$$

The within-sample reconstruction error for V_2 is:

$$RE_2=|\widehat{V_2}-V_2|$$

Then, the submatrix V_1 was projected onto the basis matrix W_2 *(dictionary),* multiplied by $W_2$and subtracted by V_1, denoting the transfer (out-of-sample) reconstruction error ($\widehat{RE_1}$) from V_2 to V_1:

$$\widehat{RE_1}=|W_2W_2^TV_1-V_1|$$

The same process for calculating $\widehat{RE_2}$:

$$\widehat{RE_2}=|W_1W_1^TV_2-V_2|$$

Finally, we calculated the absolute difference between $\widehat{RE_1}$ (i.e., the out-of-sample reconstruction error using the suboptimal W which comes from the other half of the split-sample) and RE_1 (i.e., the within-sample reconstruction error using the W optimized by OPNMF on the split-sample itself) reflecting the performance of how well a dictionary can be generalized to unseen samples:

$$DiffRE_1=|\widehat{RE_1}-RE_1|=|W_1W_1^TV_1-W_2W_2^TV_1|$$

Repeat for another split-sample (i.e., submatrix V_2):

$$DiffRE_2=|\widehat{RE}_2-RE_2|=|W_2W_2^TV_2-W_1W_1^TV_2|$$

Entries of the two *DiffRE* matrices were summed over the items and were averaged for each split-half realization denoting the out-of-sample increased reconstruction error metric.

Methodological notes for bootstrapping and 10-fold cross-validation based evaluations

In bootstrap and 10-fold cross-validation, we implemented the evaluations in a similar way as that in split-half analysis, except for the following highlighted differences:

In bootstrap (within-sample):

The four evaluation indices were calculated based on the comparison of the basis matrix *W* from the bootstrapped sample to the one derived from the original sample; the transfer (out-of-sample increased) reconstruction error was derived as follows: after projecting the left-out sample (patients that were not selected in the bootstraps) onto the dictionary obtained from the bootstrapped sample, we compared the ensuing (out-of-sample) reconstruction error with the within-sample reconstruction error of the left-out data.

In 10-fold:

We created ten partitions of equal length sampling randomly from the original sample, and performed OPNMF on nine of the ten partitions (training set), as well as the held-out one (test set).Then, we calculated the evaluation indices of aRI, JI, VI and CI between the basis matrix from the nine partitions and the one from the test sample. For transfer reconstruction error calculation, we likewise projected the held-out (1/10th) sample onto the dictionary obtained from the other nine partitions (9/10th) and compared the ensuing (out-of-sample) reconstruction error with the within-sample reconstruction error of the test set. The above process was repeated for ten times to ensure that each of the partition has been treated as the held-out (test) sample once. Afterwards, the obtained values were averaged over the ten repeats as the metric for one 10-fold realization. Finally, the above procedures were iterated for 1000 times, i.e. 1000 sets of randomly generated ten partitions.

In between-sample (PHAMOUS vs. international) bootstrap-based comparisons:

In each realization of the between-sample bootstrap-based comparison, we bootstrapped the two datasets independently and performed OPNMF separately on the ensuing bootstrapped samples. The evaluation indices of aRI, JI, VI and CI were calculated between the basis matrices derived from the two bootstrapped samples. To calculate the transfer (out-of-sample increased) reconstruction error, we projected the bootstrapped sample that was drawn from the international data onto the dictionary obtained from the bootstrapped PHAMOUS sample, and then we compared the ensuing (out-of-sample) reconstruction error with the within-sample reconstruction error of the bootstrapped international sample. This reflects how well the PHAMOUS dictionary can decode the international data, indicating the generalization performance of the PAHMOUS derived factor-structure to the heterogeneous, international patient cohorts.

Cross-sample analysis after accounting for sample size, age and illness duration differences

Because the PHAMOUS sample is larger than the international sample and the patients are older and more chronic in PHAMOUS, we performed several additional between-sample bootstrap based comparison analyses to account for these effects.

Imbalanced sample size:

We randomly subsampled 490 patients from the PHAMOUS data, and re-performed the between-sample bootstrap based comparison analysis on the ensuing PHAMOUS subsamples and the original international sample (i.e., 490 patients), repeated 5,000 times, in order to test whether an optimal four-factor solution holds when the sample is equally-sized between the two datasets. Results showed that a four-factor model remains the optimal solution in terms of stability and generalizability when sample size is balanced (Figure S8A).

Differences in age and illness duration:

We ranked the patients based on their sums of age and illness duration for each of the two samples in descending order. The patients ranked in the top 60% (for the PHAMOUS sample as the patients herein are older and more chronic) and bottom 60% (for the international sample) were included for

subsequent between-sample bootstrap based comparison analysis. Then, of the upper and lower 60% of the two samples, 80% were subsampled, and a *Wilcoxon ranksum* test was used to check if there were significant differences in age and illness duration between the subsamples. If between-subsample comparisons in age and illness duration were both non-significant (p>.05), the following evaluation steps (i.e., the "between-sample bootstrap based comparison analysis"), for assessing the stability and generalizability of PHAMOUS generated dictionaries, were continued on the newly matched datasets. Otherwise, we re-ran the 80% subsampling procedure. The whole process was repeated until 5,000 times, the 80% subsamples did not significantly differed in age and illness duration (i.e., 5,000 successful evaluations).

Results showed that a four-factor model remains the optimal solution in generalizing to the multi-site, heterogeneous international sample (Figure S8B), which is stable irrespective of age and illness duration.

Evaluations on the pooled sample (PHAMOUS + International)

To further corroborate a unified optimal factor solution, we performed additional 10-fold cross-validation and out-of-sample replication analysis after pooling the two datasets (totally 2035 patients; i.e., PHAMOUS and international). Basically, the whole pooled sample was divided into one held-out sample (20%) for replication analysis, and on the remaining 80% sample (1628 patients) 10-fold cross-validation was conducted. Afterwards, the factor models, trained based on the 80% sample, were tested on the 20% replication sample, denoting the out-of-sample generalization performance. The whole process was repeated for 5,000 times.

Results consistently showed that a four-factor model optimally captured the symptom dimensions of schizophrenia patients, in both 10-fold cross-validations (Figure S8C) and out-of-sample generalization analyses on the replication sample (Figure S8D).

Evaluation of loading and item-score predictions

The stability and accuracy of loading and item-score predictions were evaluated based on between-sample bootstrap resampling. Basically, we showed the variations of the PHAMOUS dictionary predicted factor-loadings and PANSS item-scores for the international sample over the 5,000 bootstrap

realizations, and then aggregated over subjects. In detail, in each bootstrap realization, we computed the factor-loadings for each individual patient in the international sample following the projection of international data onto the PHAMOUS generated dictionaries. Then, the predicted loadings were compared with the loadings that were estimated within the international sample. After multiplying the predicted loadings by the PHAMOUS dictionary, we got the predicted PANSS item-scores for each individual patient in the international sample. Here, the predicted item-scores were compared with the actual ratings. We used two metrics, Pearson correlation coefficient and the normalized root-mean-square-error (nRMSE), to quantify the aforementioned comparisons, denoting the precise patterns of the predicted loadings and the item-scores. Results showed that the best prediction of loadings and item-scores (averaged over the factors and the subjects) was achieved by a model with four factors where the correlation coefficient reaches highest and the nRMSE reaches lowest (Figures S10A, S10C), providing a solid support for future actionable use of the current OPNMF four-factor model. More specifically, for the four dimension loadings, prediction for the negative loadings was the most stable and accurate (Figure S10B). We also tested the prediction performance for each individual site in the international sample. Basically, the prediction accuracies for both the loadings and the item-scores for each of the nine sites were similar, and local minimums were achieved consistently across sites at a solution with four factors, indicating that the predictions from a four-factor model were stable across sites (Figure S11). In addition, the newly emerged fifth factor in the five-factor model showed the worst out-of-sample prediction accuracy (highest nRMSE and lowest correlation coefficients) with lower stability compared to the other factors in the five-factor model (Figure S10B) and the all factors in the four-factor model (Figure S10B). These results further corroborated that a four-factor model outperforms a model with five factors.

Outlier Detection

Given that the psychopathological features for some patients cannot be well captured by the OPNMF factor-models, we conducted an outlier detection step in order to filter out the patients whose PANSS scores are badly factorized by the method (i.e. the reconstruction errors are “extremely” high). First, PANSS scores of the initial assessment for all of the 1545 patients in the PHAMOUS sample were factorized by OPNMF with the factor numbers ranging from 2 to 11. Then we reconstructed the original

data matrix from the low-rank approximation by multiplying the loading matrix W^TV to the basis matrix W and computed the reconstruction error ($|V\text{-}WW^TV|$). Afterwards, we calculated three metrics (m) based on the reconstruction error matrix as follows:

1). the absolute differences between the summed values over the items for the approximation WW^TV (*S1*) and the summed values over the items for the original data matrix V (*S2*);
2). sum of the absolute values of the reconstruction error matrix ($|V\text{-}WW^TV|$) over items (for per subject);
3). a relative reconstruction error $|S1\text{-}S2|/S2$, which indicates the proportion of the absolute difference between the approximated $W*W^T*V$ and the original data matrix V relatively to V.

After the three metrics were obtained, we employed a method called median absolute deviation (MAD) rather than the standard deviation from the mean method to detect and filter out outliers, i.e. which patients are not well captured by the OPNMF factors with high reconstruction errors. This MAD method has the advantages of being insensitive to outliers (extreme values) and askew distributed data when compared to the standard deviation from the mean method (the mean and standard deviation themselves are strongly impacted by outliers) (32). The MAD method has been widely used in recent biomedical studies (33,34). The equation for calculating the *MAD* value is given as follows:

$$MAD = bM_i\left(\left|x_i - M_j\left(x_j\right)\right|\right)$$

where the x_j is the n values (here 1545) for each of the aforementioned three metrics and M_j is their median. $\left|x_i - M_j\left(x_j\right)\right|$ gives the absolute deviations from the data median, and M_i is the median of the absolute deviations. b is a constant, here 1.4826, linking to the assumption of normality of the data, regardless of the abnormality induced by outliers (35). The cutoff value (C_{mk}) for a given factor number k and a metric m was derived as follows:

$$C_{mk} = M_{mk} + 3 \cdot MAD_{mk}$$

M_{mk} is the median of the values for a metric for a given factor k, MAD_{mk} is the *MAD* for a given k and m

(here 3 times MAD was used to conservatively filter out outlier patients, which can be replaced by 2 and 2.5 which respectively refers to poorly conservative and moderate conservative as that proposed by Miller [36]). After the cutoff value was generated for each of the three metrics, patients with metric values exceeding any of the three cutoff values in more than half of the examined factor numbers (here>*5*) were rejected. The above outlier detection procedures were repeated on the whole international sample. In this stage, 14 patients were excluded for the PHAMOUS sample and 4 patients were excluded for the international sample. All of the aforementioned evaluation procedures were repeated after removal of these outlier patients.

Relationship among variables and factor analysis comparison

Inter-correlations of OPNMF factor-loadings

After controlling for overall symptom severity (i.e., total PANSS score), the inter-correlation results were differed from that without controlling for symptom severity as demonstrated in the maintext, revealing multiple anti-correlations with negative and positive loadings the lowest partial correlation coefficient (*rho*=-0.59, averaged over 10,000 bootstrap realizations; in each bootstrap realization, the original sample was drawn with replacement to form a bootstrap sample on which the partial correlation analysis was repeated), followed by the correlations between the affective and cognitive loadings (*rho*=-0.56), and the negative and cognitive loadings (*rho*=-0.41) (see Figure S15A for details).

Additional ANOVA analyses

Bootstrap and leave-one-site-out replications for 4-way ANOVAs on OPNMF factor-loadings in the international sample with controlling for total PANSS score

A bootstrap resampling procedure was performed to evaluate the stability of ANOVA findings. Specifically, in each bootstrap realization, bootstrap sample was drawn with replacement from the original (all) patients with complete age, gender and illness duration information in the international dataset and the same 4-way ANOVA analysis was conducted on the bootstrapped sample. The bootstrap stability analysis was repeated for 10,000 times. Both of the mean and the median *p*-values, as well as the *β*-values were recorded. Bootstrap results for the ANOVA analysis showed that the median *p*-values support all the significant findings in the main analysis (i.e., below 0.05).In leave-one-site-out replications, we observed that the associations between symptom severity and the four factors were all significant no matter which site was left out, while the association between age and cognitive factor was significant in five out of the eight leave-one-site-out analyses.

ANOVAs on OPNMF factor-loadings in the international sample without controlling for total PANSS score

A MANOVA was also performed to test the general effects of age, gender and illness duration on the entire set of factor-loadings without adding total PANSS score as a covariate (i.e., overall symptom severity retained). Results showed that none of the three explanatory variables affected significantly on

the joint factor-loadings (all p>.05). Then, four 3-way ANOVAs were used to explore the associations between the explanatory variables and the loadings for each factor. A significant negative association was observed between age and the cognitive factor (p=.046, β=-0.026). Positive associations of illness duration with all of the four factors were detected (negative: p=.044, β=0.041; positive: p=.017, β=0.042; affective: p=.0025, β=0.056; cognitive: p<.001, β=0.055) (Figure S15C). No significant effect of gender on any of the four factors was found. Bootstrap analysis was also performed to evaluate the robustness of the ANOVA results (repeated 10,000 times). The mean p-values of the all bootstrap experiments showed that the significant findings for the associations between illness duration and the four factors in the main analysis held. For the median p-values, all the significant findings revealed in the main ANOVA analysis held (Figure S15B).

Leave-one-site-out validation was further conducted for the ANOVAs without controlling for symptom severity. Results showed that the positive association between illness duration and the cognitive factor was significant in all of the eight leave-one-site-out analyses. The positive effect of illness duration on the positive and affective factors was significant in seven out of the eight analyses, while the other two associations (i.e., between age and the cognitive factor, and between illness duration and the positive factor) were significant or approached significant (p<.1) in six out of the eight leave-one-site-out analyses.

Four-way ANOVAs on OPNMF factor-loadings in PHAMOUS

MANOVA and 4-way ANOVAs were also performed on the PHAMOUS sample. MANOVA results showed that three of the four explanatory variables were significantly affecting the joint factor-loadings (age: p= 4.96E-7; illness duration: p=6.05E-6, total PANSS score: p=4.98E-25). Individual associations between the demographic/clinical variables and the four factor-loadings were identified by ANOVA analyses. Results showed the age to be significantly related to all four factors: negative (p=.037, β=0.011), positive (p=.015, β=-0.012), affective (p=4.50E-6, β=-0.017), and cognitive (p=1.70E-6, β=0.018), while illness duration was significantly and positively associated with positive (p=.008, β=0.014) and affective (p= 7.31E-4, β=0.013) symptoms but negatively associated with severity in negative symptoms (p= 9.15E-7, β=-0.027). Total PANSS score showed significant associations with all of the four factor-loadings (all p<4.51E-126, β>0.07).

Relationship between PANSS subscales and demographic and clinical variables

An analysis by individually putting the general psychopathology subscale (GPS), as well as the negative subscale and the positive subscale in the left hand of the ANOVA models as response variables, and age, gender, illness duration and total PANSS score in the right hand as regressors was also performed to explore the associations of the original PANSS subscales with the demographic and clinical variables in the international sample. We found that all of the three subscales were positively associated with total PANSS score (p=1.36E-84, β=0.21 for the positive subscale; p=4.14E-121, β=0.28 for the negative subscale; p=1.76E-228, β=0.51 for GPS), but no significant associations were observed for age, gender and illness duration variables. Then the ANOVA analysis was repeated after leaving the total PANSS score out of the models. The results were changed with the all three PANSS subscales showing significant associations with illness duration (positive subscale: p=.011, β=0.10; negative subscale: p=.007, β=0.13; GPS: p=.001, β=0.27).

Inter-item correlations, and correlations between OPNMF factors and PANSS subscales

The 30 individual PANSS items were correlated to each other with and without adjusting for symptom severity (total PANSS score) for the current OPNMF four-factor representation of psychopathology and the original PANSS subscales in the international sample. Results showed that the inter-item correlations for our four-factor structure are higher and more homogeneous within each of the factors than the original PANSS subscales which showed multiple anti-correlations within each of the subscales after general symptom severity was adjusted (Figure 2A-2B in the maintext). Figure S12 shows the inter-item correlation patterns without controlling for total PANSS score for both of the OPNMF factors and the original PANSS subscales

Then, we correlated the current OPNMF derived four factor-loadings with the three PANSS subscales using Pearson correlation analysis. Results showed that the four OPNMF factors were all significantly correlated with the PANSS subscales (all r>0.45, p<.001). Specifically, the negative factor showed the highest correlation coefficient with the negative subscale of PANSS (r=0.97), the positive factor showed the highest correlation coefficient with the positive subscale of PANSS (r=0.92), and both of the affective (r=0.92) and cognitive (r=0.82) factors showed the highest correlation coefficients with the GPS of PANSS. As these correlations may be explained by general symptom severity, we repeated the correlation analysis by controlling for total PANSS score. Results showed that the correlation patterns

were changed. Firstly, the correlation between the cognitive factor and GPS was not significant (r=0.02). Instead, the cognitive factor showed a weak but significant positive correlation with the positive subscale (r=0.13, p=.008) and a negative correlation with the negative subscales (r=-0.15, p=.002). Although the correlations between the OPNMF positive factor and the positive subscale (r=0.85, p<.001), as well as between the negative factor and the negative subscale (r=0.89, p<.001) were still significantly positive, the correlations of positive factor-negative subscale (r=-0.61, p<.001), positive factor-GPS (r=-0.29, p<.001), negative factor-positive subscale (r=-0.65, p<.001), and negative factor-GPS (r=-0.29, p<.001) were all significantly negative. The affective factor was significantly positively correlated with GPS (r=0.51, p<.001), but negatively correlated with both of the negative (r=-0.20, p<.001) and positive (r=-0.23, p<.001) PANSS subscales (Figure S14).

Exploratory and confirmatory factor analysis

Finally, we conducted exploratory factor analysis (EFA) on the PHAMOUS sample to derive factor-models by setting the factor numbers from four to seven according to previous literature (these factor numbers have been reported in previous PANSS factorial studies) (37-40), and then applied confirmatory factor analysis on the international sample to test the goodness of fit of the models that derived from EFA experiments. EFA was done with SPSS version 19.0 (IBM, NY, USA), and the Analysis of Moment Structures (AMOS; version 25) was used for performing CFA. Specifically, in EFA, principal components factor analysis using varimax rotation was conducted. The varimax rotation was chosen to keep consistency with previous factorial studies on PANSS (41,42). Following the varimax rotation, items were assigned to factors according to their highest loadings. Internal consistency for each of the factors was quantified by the Cronbach's alpha coefficient (higher values indicate more closely related items within a set) (43). Of note, the internal consistency analysis, as well as the below CFA was conducted in the international sample. In all CFA, PANSS items were specified to load on a single factor based on the PHAMOUS-derived EFA models. All factors were allowed to correlate with error covariates set to zero. The robust maximum likelihood method was employed to compute the fit indices, since this method is less likely to be affected by sample size, nonnormality and model size (44,45). In compliance with previous PANSS factorial studies (37-40), three indices of the Comparative Fit Index (CFI), the Normed Fit Index (NFI), and the Root Mean Square Errors of Approximation (RMSEA) were adopted to assess the goodness-of-fit. Models assessed by CFA with values of CFI and NFI greater than 0.90 and RMSEA less

than 0.08 are indicated to have adequate fit (46,47).

Results showed that the internal consistency coefficients for all of the four factor-models that identified by EFA were variable (ranging from 0.49 to 0.91) with multiple were lower than the least acceptable level of 0.7. All of these factor-models could not to be confirmed in the international sample, i.e., inadequate fit (Table S2).

Quantitative comparison of PCA and OPNMF factor models

First, we applied PCA and OPNMF to the PHAMOUS sample (as training set), then we computed the (within-sample) explained variance (EV) for the matrix reconstructed by the basis matrix (i.e., OPNMF dictionary) and the PCA loadings, respectively (Figure S13A). Of note, the higher variance explained by PCA compared to OPNMF is not unexpected as OPNMF applies a lot of regularizations/constraints which will then reduce the final variance that can be explained by the learned factors: 1) H and W (the learned parameter) must be non-negative; 2) H can be replaced by $W'V$ (projectable); 3) factors are as orthogonal as possible (W becomes sparse). Furthermore, we measured the "loss of EV" metric based on the international sample to indicate the generalization performance. As in the previous evaluations, we performed PCA and OPNMF on the international dataset, and got the within-sample EV for each of the two methods. Then, we tested how much worse (i.e., decreases in EV) when the international data were recovered by the dictionaries/components derived from the PHAMOUS sample. A higher loss of EV indicates worse generalizability of the dictionaries/components (Figure S13B). From these results, obviously that OPNMF is with better generalization performance with lower loss of EV, especially for the four factor model which achieved the local minimum. That is, PCA showed higher within-sample EV, but at the cost of much lower interpretability. In turn, OPNMF showed a slightly lower EV, but it better generalized to new data with lower loss of EV when the trained dictionaries were applied to novel samples. In summary, the good generalization combined with the superior interpretability of a parts-based representation make OPNMF a more appropriate tool for representing latent dimensions of psychopathology than PCA.

Identification of psychopathological subtypes

Methodology and cluster selection

After the effects of gender, age, illness duration and symptom severity (total PANSS score) were partialled out from the four factor-loadings in the international sample, fuzzy c-means clustering (48) was performed on the ensuing residuals to partition the patients into symptomatically distinct subgroups, which also provided probabilistic cluster memberships for each patient. The object function for fuzzy c-means contains a fuzzy partition matrix so that each subject is allowed to belong to multiple clusters with varying degree of membership. The fuzzifier *m* (i.e., the exponent for the fuzzy partition matrix [*U*]; $1 < m < \infty$) controls the amount of fuzzy overlap between clusters (how fuzzy the boundaries between clusters can be) with larger values resulting in fuzzier clusters, i.e. a greater degree of overlap. Here we used a value of 2.0 for *m* as a standard setting which was common in previous literature and has been empirically supported with good performance (49-51). We also inspected the cluster solution by switching m = 2 to m = 1.5 or 2.5. The squared Euclidean distance was used as the distance metric between subjects. The resulting membership values reflect how strong a patient is attributed to each cluster, based on which we can assign the patients to respective clusters according to the maximal membership degree. We set the cluster number *c* ranging from 2 to 9 to search for the optimal cluster solution representing the schizophrenia psychopathological subtypes. The optimal cluster number was determined based on three validity indices of the fuzzy Silhouette index (SI) (52), the Xie and Beni index (XB) (53), and partition entropy (PE) (48), testing the segregation, compactness and fuzziness of the resulting partitions. Calculations of XB and fuzzy SI both require the fuzzy partition matrix *U* (membership degree), as well as the original data points which are thus directly connecting to the geometric feature of the data. The XB index further requires the cluster centroid metric (defined as the mean of all data points within a cluster, weighted by their degree of belonging to this cluster). PE can be calculated by given only the fuzzy partition matrix *U*, reflecting the fuzziness of the cluster partitions. Higher values of fuzzy SI and lower values of XB and PE indicate better clustering quality. Detailed descriptions for these validity indices are given as follows:

1). A fuzzy extension of the original SI (54) was employed in the decision of the optimal cluster solution derived by fuzzy c-means. This fuzzy SI is an extension of the original silhouette criterion (namely the

crisp SI) to account for the data set containing overlapping clusters (52). The fuzzy SI has been shown to perform equally well to or better than other validity indices for fuzzy cluster analysis (52). The crisp SI is the mean Silhouette width of all the samples and it assesses the compactness and separation of hard cluster partitions. The silhouette width for a sample is defined as:

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$

where a_i refers to the average distance between a sample and all other samples in the same cluster they are assigned, and b_i is the minimum average distance from the sample to the all samples in other clusters. SI could have a value between -1 and 1. Higher SI corresponds to better clustering quality, i.e. each sample better lies within its cluster. In the case of fuzzy cluster analysis, the crisp SI may fail to discriminate between overlapping data clusters since it does not utilize the membership information embedded in the fuzzy partition matrix *U* on degrees to which clusters overlap one another. The equation for calculating fuzzy SI is given as follows:

$$FSI = \frac{\sum_{i=1}^{N}(\mu_{pi} - \mu_{qi})^{\alpha} s_i}{\sum_{i=1}^{N}(\mu_{pi} - \mu_{qi})^{\alpha}}$$

where s_i is the silhouette width of subject *i*, *pi* and *qi* refer to the first and second maximal elements of the *i*-th row of the fuzzy partition matrix (i.e., the subject-by-partition membership degree μ), respectively. α is a user-optional weighting coefficient; lower α approaches the crisp SI while higher α tends to uncover smaller regions with higher data densities when existing subclusters. Here, α=1 was used as the default setting by the developers (52).

2). Involving multiple validity indices to determine the optimal cluster solution is required. Another representative validity index for fuzzy cluster analysis - the XB index - was then adopted. This fuzzy clustering specific validity index was shown to be reliable (53). Calculation of the value for this index requires both the membership probabilistic matrix *U* and the cluster prototype (centroid) for each partition. This index, as the fuzzy SI, has a direct connection to the geometry of the original data. The XB index is defined as:

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{2} \left\| C_i - X_j \right\|^2}{n \min\limits_{i \neq j} \left\| C_i - C_j \right\|^2}$$

where X_j (j = 1, 2, … ,n) refers to the number of data points of a fuzzy c-partition of the data set X, C_i is the center of cluster i, μ_{ij} is the membership likelihood of data point j in cluster i. The numerator denotes the summation of the squares of fuzzy deviation for the total data points n to the centers of all cluster partitions c, reflecting the tightness of the discovered clusters. The fuzzy deviation of a data point j from cluster i is defined as the distance between this data point to the center of cluster i weighted by the membership degree of this data point pertaining to cluster i. The denominator indicates how well the fuzzy partitions are separated, and the value of which was derived by calculating the minimum distance between the centers of two neighbouring clusters. In this form, smaller values of *XB* represent for more compact and well-separated clusters.

3). Besides, a classic fuzzy cluster validity index of PE (48) was employed to complement the above two validity indices. The value of PE can be calculated given only the probabilistic membership degree information (i.e., the fuzzy partition matrix U), which reflects the fuzziness of the cluster partitions. The PE index is defined as follows:

$$PE = -\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{c} \mu_{ij} \cdot \log_{\mathrm{a}}(\mu_{ij})$$

where a$\in$(1, ∞) is the base of the logarithm [here we used a default base a of exp(1)], μ_{ij} is the membership likelihood of data point I in cluster j, n is the number of total data points (subjects), c refers to cluster number. Computation of the PE index requires c greater than 1 and its value ranges in [0, $\log_a c$]. Smaller value of PE indicates crisper clustering solution.

Assessment of clustering stability

After the optimal cluster number was determined, leave-one-site-out analysis was performed to validate that the clustering results are not driven by any particular site. Specifically, in each leave-one-site-out

experiment, we left out all the patients from one site, and on the remaining sample we re-calculated the residuals after partialling out the effects of age, gender, illness duration and symptom severity from the factor-loadings. Then, the same fuzzy c-means clustering strategy was applied to the obtained residuals followed by the same cluster selection criteria to determine the optimal number of clusters. These processes were repeated until each site had been left out once. Stability of cluster solutions was tested via subsampling and bootstrap resampling evaluation processes, reflecting how stable the partitions hold when the original sample is perturbed. If the structure in the data has been captured well by a partition, this partition should be stable with respect to data perturbation. The evaluation scheme was implemented as follows: the whole dataset is clustered by fuzzy c-means; a set of random subsamples (70% of the whole dataset) and bootstrapped samples are generated and clustered as well. ARI was used to indicate stability, which reflects the similarity between the partition of the reference clustering and the partitions of the subsampled and bootstrapped data, i.e. the consistency of patient-pair assignment between the sub/bootstrapped partitions and the partition derived from the original sample. The best partition in representing the structure of the original sample should have the highest stability (aRI). The idea and the process of testing clustering stability in the present study were similar as that demonstrated in previous literature (55). Also, values of the three employed validity indices were calculated for each subsampled and bootstrapped data, to verify whether the optimal cluster solution holds when the original data were perturbed. Of note, for any new subsampled and bootstrapped data, residuals that used for clustering were re-calculated based on the corresponding covariates of age, gender, illness duration and total PANSS score.

A fuzzy c-means 2-cluster solution was robustly identified as the best representation of schizophrenia psychopathological subgroups by switching the fuzzifier m = 2 to m = 1.5 or 2.5, as well as in leave-one-site-out analysis (Figure S17), subsampling and bootstrap resampling stability evaluation experiments (Figure S16).

Comparison of ambiguous class with core subtypes

Four factor-loadings (adjusted) for the ambiguous patient class were compared with the identified two core subtypes by 10,000 permutation tests with shuffled cluster labels. The 25% of patients with no clear cluster memberships (ambiguous class) were in the “middle” in terms of symptomatology compared to

the two core subtypes, showing close-to-zero means and high standard deviations in all of the four (adjusted) factor-loadings. Specifically, the ambiguous class (mean ± SD, -0.23 ± 1.91) showed significantly higher negative symptoms than subtype B (mean ± SD, -0.87 ± 0.88; *p*=.004), but significantly lower negative symptoms than subtype A (mean ± SD, 0.84 ± 0.98; *p*=.002). For positive symptoms, the ambiguous class (mean ± SD, -0.12 ± 1.47) was significantly lower than subtype B (mean ± SD, 1.84 ± 0.99; *p*=.002) but was significantly higher than subtype A (mean ± SD, -1.39 ± 0.78; *p*=.002). For the affective dimension, the ambiguous class (mean ± SD, 0.16 ± 1.71) was only significantly higher than subtype B (mean ± SD, -0.32 ± 0.78; *p*=.007). No significant differences were detected for the cognitive dimension between the ambiguous class and the other two clear clusters.

Longitudinal stability analysis

Fuzzy c-means clustering with the same parameter settings as the one applied to the international sample was performed on the patients with repeatedly assessed PANSS scores in the PHAMOUS sample. 527 patients who have the complete age, gender and illness duration information were involved. The optimal dictionary with four factors, identified on the initially assessed PANSS scores of 1545 patients in the PHAMOUS sample, was used for projection to yield the factor-loadings. Effects of age, gender, illness duration and symptom severity (total PANSS score) on the projected loadings were regressed out, and the residuals were used for clustering analysis to identify patient subtypes. The aforementioned three validity indices were employed to ascertain the optimal cluster number, and we found that all the values pointed to a cluster solution equaling to 2 as well. Then, we assessed the longitudinal stability of the identified psychopathological subtypes as follows (i.e., whether a patient preserved his/her subtype over time from the initially recorded PANSS scores to the follow-up psychopathology):

A. The optimal four-factor dictionary from the PHAMOUS sample based on the initially assessed PANSS scores of 1545 patients, was used as reference, on which the initially assessed PANSS scores of the 527 follow-up patients were projected to derive the factor-loadings.

B. On the factor-loadings we performed 4-way ANOVA analysis and recorded the resulting betas for the four factor-loadings.

C. Effects of age, gender, illness duration and symptom severity (total PANSS score) on the four factor-loadings were then removed to obtain the residuals. We performed fuzzy c-means on the

residuals to partition the patients into 2 clusters and those patients with membership likelihoods lower than 0.7 were excluded. The subtype (cluster label) for each patient and the cluster centers were recorded.

D. Likewise, the four-factor dictionary from the whole PHAMOUS sample was used for deriving the factor-loadings of the repeatedly assessed PANSS scores of the 527 follow-up patients. Of note, we here constructed a regression model using the betas that obtained in B to derive the residuals after removing the effects of age, gender, illness duration and symptom severity on the factor-loadings.

E. Based on the residuals obtained in D, we calculated the squared Euclidean distance between each patient and the two cluster centers defined in C (i.e., based on the initially assessed PANSS scores). Then, each patient was assigned to a specific cluster if its center, comparing to other clusters, has the closest distance to that patient. Afterwards, each patient has a "predicted" cluster label (i.e., psychopathological subtype) according to the follow-up PANSS assessments. In this stage, by comparing the cluster label of each patient based on the repeatedly assessed PANSS scores to the one identified based on the initial assessments derived in C, we have the information on how many patients retained their subtypes longitudinally.

Among the all 527 patients with repeatedly assessed PANSS scores in the PHAMOUS sample, 290 patients were retained after filtering out those with ambiguous cluster attributes (membership likelihoods μ<0.7). The longitudinal stability analysis of sub-typing showed that 78.62% (228/290) patients kept their subtype attributes. Furthermore, patients assigned as subtype B at baseline were in higher frequency converting to subtype A at follow-up (27.5%) than the patients transforming from subtype A to subtype B (14.9%). In detail, 62 patients changed their cluster memberships with 21 patients initially assigned to subtype A and 41 assigned to subtype B. Of the 21 patients that switched from subtype A at baseline, 13 patients were clearly assigned to subtype B (membership values μ>0.7) at follow-up, and the remaining ones became ambiguous (i.e., the mixed group). In the 41 patients changing from subtype B at baseline, 19 patients were clearly assigned to subtype A at follow-up with 22 in the ambiguous class due to their membership values lower than 0.7. Among the 237 ambiguous patients (μ<0.7) identified based on the initially assessed PANSS scores, 66 patients were clearly assigned to subtype A at follow up (μ>0.7) and 52 patients were clearly assigned to subtype B. The remaining 119 patients kept their ambiguous attributes (μ<0.7).

Additional clustering analyses

For a comparison with the clustering based on our OPNMF four factor-loadings with an adjustment step, we applied the same clustering methods (fuzzy c-means and Gaussian mixture modeling [GMM]) to the individual psychopathology expressed by the three original PANSS subscales as well as by the 30 single-item PANSS scores with adjusting for covariates (i.e., age, gender, illness duration and total PANSS score). In addition, we repeated these two analyses as well as the one based on the OPNMF factor-loadings without adjusting for covariates. Finally, we also performed clustering on the factor loadings without adjusting for total PANSS, i.e., symptom severity.

Clustering based on the individual 30 PANSS items and the three PANSS subscales:

Clustering the 30 single-item PANSS scores

1) Clustering the original 30 individual PANSS item-scores without adjustment for covariates:

Fuzzy c-means yielded an optimal two-cluster solution (Figure S19A; see in particular the XB index and PE). The clusters were mainly driven by overall symptom severity, such that scores for all 30 items were either high or low, respectively (Figure S19C). Similarly, the four clusters identified by GMM, based on Bayesian information criterion, were reflected to overall symptom severity (Figure S19D).

2) Clustering residuals of the 30 individual PANSS item-scores after adjusting for the four covariates:

Fuzzy c-means provided ambiguous clustering (for all patients the cluster membership degree was close to 0.5). Also, values of the internal validity index of XB are extremely large, indicating poor clustering quality (Figure S19B). GMM provided a “single cluster” solution, i.e., no evidence for subgroups.

Clustering the PANSS subscales

1) Clustering original PANSS subscales without adjusting for covariates:

Similar to item-wise clustering above, both fuzzy c-means (2 clusters; Figure S20A) and GMM (6 clusters; Figure S20B) yielded clusters that are driven by overall symptom severity.

2) Clustering residuals of the PANSS subscales after adjusting for covariates:

For fuzzy c-means, the model selection criteria were inconsistent (XB index: 6 [Figure S21A] and 7 [Figure S21E]; fuzzy SI: 3; and PE: 2). Moreover, the yielded partitions are less stable. As shown in Figure S21B-S21C, values of the aRI, which reflects the similarity of patient-to-cluster assignment between the

subsamples and bootstrapped samples, were generally decreased when comparing to the partitions generated based on our four OPNMF factors (Figure 3B & Figure S16).As there is no consistently selected number of clusters, we set the cluster number to 2 for a direct comparison with the clusters generated based on our four factors (Figure S21D). Results showed a moderate agreement between the partitions generated based on the PANSS subscales and our four factors (aRI value of 0.65). GMM again provided a "single cluster" solution (selected using Bayesian information criterion), i.e., did not reveal any subgroups.

Overall, when not adjusting for age, gender, illness duration, and total PANSS score, we get the severity clustering (as for the clustering on the raw factor-loadings). When we do adjust, we get an "ambiguous cluster" solution (fuzzy c-means) or a "single cluster" solution (GMM).

Clustering based on the 4 factor loadings from the current study, either without adjustment for (any) covariates or without adjusting for disease severity (total PANSS score):

Clustering the raw factor-loadings on the four OPNMF factors without covariates adjustment

Fuzzy c-means yielded a two-cluster solution, optimally representing the patient subgroups in schizophrenia (Figure S22A), which is featured by either high or low overall symptomatology (Figure S22D-S22E). Moreover, stability of the resulting partitions was much lower (Figure S22B-S22C) than after adjustment done in the main analyses. GMM yielded six clusters which are mainly driven by overall symptom severity (Figure S22F).

Clustering based on our four factors with age, gender and illness duration adjustment, but without adjustment of total PANSS score:

Results of fuzzy c-means showed an optimal two-cluster solution (Figure S23A-S23B). One of the clusters is prominent in all of the four symptom dimensions, while the other cluster is significantly lower in these dimensions (low symptomatology) (Figure S23C-S23D). GMM, likewise, yielded three severity clusters (Figure S23E).

Overall, clustering without any covariates or symptom severity adjustment yields clusters driven by overall symptom burden. These new analyses thus indicate that clustering with such an adjustment step is necessary when trying to identify latent subgroups among schizophrenia patients. Otherwise, clusters

will primarily reflect overall disease severity.

In summary, results of the additional clustering analyses clearly show that only the four-factor model with adjustment for all covariates provides a clear, clinically meaningful clustering of patients into subgroups.

Classification of psychopathological subtypes based on resting-state functional connectivity

MRI data acquisition and preprocessing

A high-resolution T1-weighted anatomical scan and a resting-state fMRI scan were obtained for each of the 147 patients in the international sample. Specific scanning parameters varied by site, please refer to Table S3 for detailed information.

Image preprocessing was done with the various programs in Statistical Parametric Mapping software (SPM12; https://www.fil.ion.ucl.ac.uk/spm) and Computational Anatomy Toolbox (CAT12; https://www.neuro.uni-jena.de/cat). In detail, for resting-state modality, the first four volumes from all fMRI scans were removed to allow for MR signal equilibrium and adaptation of subjects. Prior to other preprocessing processes, we first estimated the head motion parameters to avoid a potential underestimate of movement if the slice-time correction step is performed first. Here we employed a metric of DVARS (D referring to temporal derivative of time-courses, VARS referring to root mean square [RMS] variance over voxels across the whole brain) (58) as an evaluator of head motion by calculating the voxel-wise BOLD signal intensity change between one frame (timepoint) and its backward to detect and remove those patients with excessive movements. The equation for calculating DVARS was given as below:

$$DVARS(\Delta I_i) = \sqrt{\langle [\Delta I_i(\vec{x})]^2 \rangle} = \sqrt{\langle I_i(\vec{x}) - I_{i-1}(\vec{x})]^2 \rangle}$$

where $I_i(\vec{x})$ is image BOLD signal intensity at locus $\vec{x}$on frame *i*and angle brackets denote the spatial average of the voxel-wise signal intensity changes over the whole brain. Finally, the DVARS metric was scaled by dividing by the median brain intensity and then multiplying by 1000 to approximate the magnitude that was reported in Power *et al.* (58), i.e., 10 units of DVARS refer to 1% BOLD signal change. This is a critical step before any subsequently quantitative investigations since excessive head motion will lead to spurious signals that bias the functional connectivity measures (58,59).

According to the histograms of the mean DVARS values for the patients (Figure S24), any patient with a DVARS larger than 50 (i.e., 5% BOLD signal change) was treated as an outlier and was removed

from the subsequent analyses. Among all the 96 patients (after filtering out five subjects that were identified with an artifact in their T1 segments in the following quality control step based on structural images) of core subtypes with resting-state fMRI data in the international sample, 12 patients (4 for subtype A and 8 for subtype B) were identified with high head motions (mean DVARS> 50) and were thus excluded for the classification analysis. This retained totally 84 patients [age (33.55 ± 11.28 years), gender (male/female: 59/25), illness duration (9.90 ± 9.65 years), and total PANSS score (58.05 ± 17.69)] with 47 in subtype A and 37 in subtype B. Between-subtype comparison of DVARS was achieved by using the *Wilcoxon rank-sum* test. This test showed no significant between-subtype difference (p=.10) with respect to the DVARS metric. Of note, this threshold of DVARS = 50 is roughly equivalent to a framewise displacement (FD) of 0.5 mm as demonstrated in Power *et al.* (58), and the cutoff of FD = 0.5mm has been commonly used in recent studies (as reviewed in Power *et al.*[59]). Then, all of the images were slice timing corrected using a newly proposed method of filter shift (60). The effectiveness and superiority of this method over the existing interpolation-based methods have been elaborated (60), especially in the case that the subjects have moderate to high head motions. Specifically, the signal value between the sample points can be more precisely estimated by using the Kaiser multiplicative windowed *Sinc* filtering than the *Sinc* interpolation used in SPM and the Hanning interpolation (a smoothed window function as to attenuate the rippling artifact encountered in the *Sinc* interpolation) implemented in FMRIB software library (FSL) (61). The slice timing corrected echo-planar imaging (EPI) volumes were then head motion corrected using the realignment program in SPM12, and the derived six motion parameters (three translations, three rotations) were saved and used as regressors for subsequent statistical analyses. Following head motion correction, the EPI data were normalized to MNI152 space by using an EPI template in SPM12 (affine transformation of each subject's EPI data to this EPI template followed by nonlinear registration to the EPI template) with a 4 x 5 x 4 basis set to alleviate overfitting. This method has been recently proposed when distortion correction cannot be conducted in the condition that the field map is not available, which showed improved spatial convergence between EPI and structural images and within-dataset similarity (62). The normalized EPI images were resampled to an isotropic voxel size of 2mm. The high-resolution T1-weighted structural images were also preprocessed, including tissue segmentation and spatial normalization to MNI152 space which we used the shoot program (63) in CAT12. Then, the partial volume image for each patient, which contains all of the three grey matter (GM), white matter (WM) and CSF tissue types, was used as masks for extracting

the mean WM and CSF signals. To avoid an inaccurate estimate of the non-neuronal confounding signals, a quality control step was conducted to filter out those patients whose segments of T1 images had apparent artifacts and/or poor segmentation quality by using the "check sample homogeneity" module in CAT12. In this stage, five patients were filtered out, and the number of MRI images reported for classification analysis (i.e., 84 patients as aforementioned) is the number after this filtering step. Confounders of 24 head motion parameters (the 6 head motion parameters of roll, pitch, yaw, translation in three dimensions, their first temporal derivatives, and quadratic term signals), together with the non-neuronal components of the extracted total WM and CSF signals were regressed out from the overall BOLD signals (64). Of note, we also generated a version of preprocessed images with additionally global mean signal regression (GSR) as recent studies have suggested an improvement in brain prediction for behavior ratings with GSR (65), though this step remains a controversial option in resting-state fMRI preprocessing (66,67). Finally, band-pass filtering was performed on the data to restrict frequencies between .01 and .08 Hz.

Connectivity matrix construction and classification analysis

The preprocessed EPI images were used to generate a whole-brain connectome for each patient, quantifying functional connectivity between regions of interest spanning the entire brain in terms of correlated, spontaneous fluctuations in the resting-state BOLD signal. In the present study, time-series were extracted based on a parcellation system after combining Schäfer's 600 cortical parcels (local-global parcellation based on resting-state functional connectivity) (68) with the Brainnetome 36 subcortical parcels (69). The used cortical parcels were demonstrated to agree with the boundaries of certain cortical areas defined using histology and visuotopic fMRI, revealing neurobiologically meaningful features of brain organization. The Brainnetome atlas is fine-grained and cross-validated, containing information on both anatomical and functional connections. Of note, although Schäfer's parcellation system has been shown to perform better than other atlases (70-73), in particular for resting-state analyses, it doesn't currently feature a subcortical parcellation, so the Brainnetome subcortical parcels - which are derived based on diffusion tractography - were added. For cortical parcellation, resting-state modality should be preferred, as long-distance tracts, as well as those targeting heterotopic contralateral areas, are hard to resolve using diffusion weighted imaging (DWI). For subcortical regions, in turn, where

the fibers fan out but resting-state scanning is probably too coarse in resolution, DWI is optimal. So basically, we combined the best of both worlds. Nevertheless, we performed a control analysis after replacing the Brainnetome 36 subcortical parcels by a unified resting-state functional connectivity derived subcortical parcellation, i.e., Yeo's 7 network striatum extension (74). Schäfer's cortical parcellation has been shown to largely preserve Yeo's network structure (68). The extracted voxel-wise time-series for each of the 636 parcels were compressed using the first eigenvariate which were then used to calculate pairwise Pearson correlations to form the whole brain connectivity matrix. The correlations were finally *Fisher's* z-transformed prior to classification analysis. This parcel-based approach has the merits of enhanced computational tractability and biological interpretability (avoided the massive whole brain voxels while also extended the conventional seed-based approaches). A supervised support vector machine (SVM) was adopted to approach the classification problem, to classify the psychopathological subtypes for novel patients from the resting-state fMRI features. SVM learns the relationship between a set of input variables or features, and a particular outcome across a set of observations. The goal of SVM is to fit a function which approximates the relation between the features and the outcomes that can be used later to infer the outcomes for a new observation given its features. In the present study, we used a non-linear extension of SVM, namely the Radial Basis Function (RBF) kernel SVM, which could accommodate the potential non-linear relationship between the neural space and psychopathology. Parcel-wise classification analysis was conducted using the connectivity profile of each parcel individually. Effects of age, gender, site, illness duration, symptom severity (total PANSS score) and head motion parameter on the connectivity matrix in the test sample were adjusted using the betas fitted only in the training sample (75). A nested 10-fold grid-search was implemented among (only) the training data to tune the hyperparameters of C (the error/margin trade-off parameter; in SVM classification case, its target function attempts to find a separating hyperplane based on the feature space that is minimizing a measure of error on the training set while simultaneously maximizing the 'margin' between the two classes) and γ (the kernel parameter) for the RBF kernel. Sample imbalance was addressed by setting class weights when training the RBF-SVM models, as well as a stratified 10-fold cross-validation strategy for assessing the out-of-sample classification performance. The resulting balanced-accuracy for each parcel was averaged over folds and then over 50 replications of the entire procedure to avoid influences of the initial splits. The balanced accuracy (76) was calculated as follows:

$$Balanced\ accuracy = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N})$$

Where *TP* is the number of true positives: here the number of patients correctly classified to subtype B; *TN* is the number of true negatives: here the number of patients correctly classified to subtype A; *P* is the total number of patients with subtype B for classification; *N* is the total number of patients with subtype A for classification.

Finally, permutation tests by shuffling the psychopathological cluster (subtype) labels were implemented to assess whether the classification accuracy for each of the parcels was significantly above chance (77, 78). Specifically, in the realization of each permutation test: (1) the original cluster labels identified at psychopathological level were randomly permuted to obtain a randomized version of labels, (2) the same nested 10-fold cross-validation was conducted on all of the parcels using a set of uniformly permutated labels, and the balanced classification accuracy was recorded for each parcel. Then, the above two steps were repeated for 250 times. The whole process allowed us to create an empirical distribution of the classification accuracy to compare the (true) accuracy of each parcel against. A parcel with an accuracy rate of using the original (non-permutated) cluster labels above the 95th percentile of this distribution was identified as statistically significant (p<.05) in classifying the schizophrenia psychopathological subtypes (if the accuracy rate of non-permutated cluster labels exceeds those obtained from the all 250 permutation tests, indicating a statistical significance of p=.004). We furthermore applied a false discover rate (FDR) correction procedure to address the issue of multiple comparisons across the all parcels to avoid false-positive classifiable parcels (FDR q<.05 corrected). Classification with GSR showed that 48 parcels were identified as significantly classifiable for the psychopathological subtypes after FDR correction for multiple comparisons. Patterns of significant classifiable parcels were generally stable across the analyses with and without GSR (Figure 4B & Figure S26A). Top classifiable brain regions (e.g., right temporoparietal junction, ventral medial prefrontal cortex, as well as bilateral precuneus and posterior cingulate cortex) were replicated, and the highest classification accuracy was actually slightly increased for the GSR version (73%) compared to the non-GSR version (70%). Also, classification (FDR corrected) after replacing the Brainnetome 36 subcortical parcels by a functional subcortical parcellation of Yeo's 7 network striatum extension largely

replicated the main results (Figure S26B).

Functional characterization analysis

After the classification analysis was done, a functional decoding analysis was performed on the significant classifiable brain parcels to characterize their functions based on the BrainMap database (http://brainmap.org/) which recruited the "behavioral domain" and "paradigm class" meta-data of prior neuroimaging experiments (79). Five main categories "cognition, action, perception, emotion, interoception," as well as their respective subcategories constitute as the behavioral domains. The specific tasks conducted in the respective experiment are categorized into various paradigm classes. To depict the individual functional profile of each significant cluster, we employed quantitative "forward inference" and "reverse inference" experiments as previously described (80-82). Briefly, forward inference assesses a cluster's functional profile by identifying taxonomic labels for which the probability of finding activation in a specific cluster is significantly higher than the chance of finding activation for that particular cluster across the entire database. Statistical significance of the forward inference was estimated using a binomial test followed by an FDR correction at the level of $p<.05$ (81,83). In contrast, the reverse inference approach determines a cluster's functional profile based on Bayes' rule by means of identifying the most likely behavioral domains and paradigm classes given activation in a particular cluster. A $\chi 2$ test was employed to establish the significance of reverse inference using the same FDR correction strategy ($p<.05$) to account for multiple comparisons. To ensure robustness, results were shown for those of congruent forward and reverse inferences. The characterized functions for the all 53 significant classifiable parcels that survived from FDR corrections were presented in Table S4.

Supplemental Tables

Table S1.Demographic and clinical characteristics of the schizophrenia samples for each site

Characteristics	Europe						The USA		Asia	Total
	Aachen	Lille	Göttingen	Groningen	Utrecht	Munich	Albuquerque	Wayne State	Singapore	
Total *N* for OPNMF	89	53	35	28	50	21	48	19	147	490
Age	35.73±9.59	35.09±9.45	32.31±9.83	35.25±11.19	29.98±9.68	34.05±12.27	37.67±13.76	30.53±8.62	32.72±9.10	34.89±11.67
Gender (male/female)	57/32	41/12	28/7	17/11	32/18	10/11	37/11	10/9	101/46	333/157
N with illness duration information for ANOVA and sub-typing	70	40	35	23	11	0	48	19	147	393
Illness duration	8.90± 8.39	12.60± 6.81	7.29±7.72	8.39±7.94	9.00± 7.06	N.A	16.69± 12.47	7.79± 8.29	6.55±7.47	9.13±8.98
PANSS										
P3 item	2.16±1.73	5.13±0.83	4.07±1.71	1.46±1.02	3.60±1.99	2.57±1.81	2.69±1.49	1.95±0.94	1.88±1.25	2.66±1.83
Positive	14.54±5.77	21.32± 4.76	15.82± 5.06	12.03±3.54	16.14±4.43	19.38±6.15	14.50±4.97	11.63±2.16	10.59±3.84	14.24±5.76
Negative	19.38±8.90	21.02±6.08	14.21±4.64	13.57±4.87	16.28±4.99	20.90± 7.75	14.15±4.48	11.63±3.25	9.00±3.03	14.67±7.21
General	34.64±13.68	40.85±10.65	29.11±7.74	27.97±5.86	33.20±7.68	39.95 ±11.09	28.21±8.36	20.79±3.79	20.21±3.70	29.10±11.34
Total	68.56±24.69	83.19±19.06	59.14±14.92	53.57±11.04	65.62±14.32	80.24±21.70	56.85±13.52	44.05±7.55	39.80±8.39	58.01±21.87

Characteristics	Europe						The USA		Asia	Total
	Aachen	Lille	Göttingen	Groningen	Utrecht	Munich	Albuquerque	Wayne State	Singapore	
N with medication information	29	17	38	24	15	N.A	69	19	N.A	211
Antipsychotic treatment										
FGA	0	3	0	1	5	N.A	8	9	N.A	26
SGA	27	12	29	23	9	N.A	58	9	N.A	167
FGA + SGA	2	2	7	0	1	N.A	3	1	N.A	16
None	0	0	2	0	0	N.A	0	0	N.A	2
OZP-equivalent dosage	20.64±11.03	26.24±21.21	25.06±11.49	14.55±8.31	17.10±12.42	N.A	14.84±10.96	12.47±12.88	N.A	19.64±14.15

Note: Data are mean±SD; N: number of subjects, OPNMF: orthonormal projective nonnegative matrix factorization, FGA: first-generation antipsychotic, SGA: second-generation antipsychotic, PANSS: Positive and Negative Syndrome Scale, OZP: Olanzapine, NA: not available.

Table S2. **Fit indices for confirming the PHAMOUS sample derived four to seven factor models on the international sample**

Models	NFI	CFI	RMSEA	Cronbach's alpha
Four-factor	0.712	0.756	0.099	0.59, 0.82, 0.82, 0.91
Five-factor	0.757	0.804	0.089	0.59, 0.79, 0.81, 0.83, 0.91
Six-factor	0.748	0.793	0.092	0.49, 0.59, 0.78, 0.79, 0.81, 0.91
Seven-factor	0.756	0.800	0.091	0.49, 0.59, 0.69, 0.70, 0.78, 0.79, 0.91

Note: NFI: Normed Fit Index, CFI: Comparative Fit Index, RMSEA: the Root Mean Square Errors of Approximation.

Table S3A. Scanning parameters for resting-state BOLD fMRI across study sites

Site	Aachen-1	Aachen-2	Groningen	Göttigen	Lille	Albuquerque	Utrecht
Scanner	Siemens TrioTim3T	Siemens TrioTim3T	Philips Achieva3T	Siemens TrioTim 3T	Philips[1] Achieva 3T	Siemens TrioTim 3T	Philips[1] Achieva 3T
TR (ms)	2000	2000	2400	2000	19.25	2000	21.75
TE (ms)	21	28	28	30	9.6	29	32.4
Number of slices	44	34	43	33	45	32	40
Slice-thickness (mm)	3	3.3	3	3	3.22	4	4
Gap (mm)	n.a	3.6	n.a	0.6	n.a	1	n.a
FA (degree)	n.a	77	85	70	9	75	10
Orientation	Axial	Axial	Axial	Axial	Sagittal	Axial	coronal
In-plane resolution (mm^2)	3 x 3	3.6 x 3.6	3.44 x 3.44	3 x 3	n.a	3 x 3	n.a
Voxel size (mm^3)	3 x 3x 3	3.6 x 3.6x 3.3	3.44 x 3.44x 3	3 x 3x 3	3.22 x 3.22 x 3.4	3 x 3x 4	4 x 4 x 4

Note: TR: repetition time, TE: echo time, FA: flip angle; [1]PRESTO-SENSE sequence achieving full brain coverage within 609ms for the Utrecht site and 1001ms for the Lille site combining a 3D-PRESTO pulse sequence with parallel imaging in 2 directions (8-channel SENSE head-coil).

Table S3B. Scanning parameters for T1-weighted structural MRI across study sites

Site	Aachen-1	Aachen-2	Groningen	Göttigen	Lille	Albuquerque	Utrecht
Scanner	Siemens TrioTim3T	Siemens TrioTim3T	Philips Achieva 3T	Siemens TrioTim3T	Philips[1] Achieva 3T	Siemens[2] TrioTim3T	Philips[1] Achieva 3T
TR (ms)	1900	2300	2500	2250	10	2530	9.86
TE (ms)	2	3.03	4.6	3.26	4.6	[1.64, 3.5, 5.36, 7.22, 9.08]	4.6
Number of slices	176	176	160	176	160	176	160
Slice-thickness (mm)	1	1	1	1	1	1	1
FA	n.a	9	30	n.a	n.a	7	n.a
In-plane resolution (mm^2)	0.97 x 0.97	1 x 1	1 x 1	1 x 1	1 x 1	1 x 1	0.875 x 0.875

Note: TR: repetition time, TE: echo time, FA: flip angle; [1]PRESTO-SENSE sequence, [2]a multi-echo MPRAGE (MEMPR) sequence with 5 TEs.

Table S4. Classification accuracies and functional characterizations for the 53 parcels showing significant classification performance in discriminating the two psychopathological subtypes among the whole 636 parcels (FDR corrected $p < .05$)

Parcel	Side	Anatomical region	Balanced accuracy	Cognitive domain	Paradigm classes
1	R	vmPFC	0.703	Emotion.fear/reward; cognition.reasoning	Reward, face monitor/discrimination; Gambling
2	R	TPJ	0.688	Social cognition	Theory of mind
3	L	PCC	0.679	Cognition.memory explicit	*n.s.*
4	R	vmPFC	0.677	Emotion.reward	Reward; face monitor /discrimination
5	L	precuneus	0.674	Cognition(Language, social cognition, memory explicit); emotion	Theory of mind; emotion induction
6	R	precuneus	0.670	Cognition.memory explicit	Episodic recall
7	R	precuneus	0.667	Action.preparation	*n.s.*
8	L	lingual gyrus	0.665	Cognition.memory explicit	*n.s.*
9	L	Heschl's gyrus/insular cortex	0.657	Action.execution speech; Perception.Audition; cognition.music; perception Somesthesis.Pain	Music production; Passive listening; Recitation /repetition (overt); Pain monitor /discrimination
10	L	vmPFC (frontal medial cortex)	0.649	Emotion.reward; cognition. social cognition	Theory of mind; Reward; Taste
11	L	precuneus	0.646	Cognition.social cognition	Theory of mind
12	L	superior occipital cortex	0.646	Cognition.memory explicit	Cued explicit recognition/recall
13	R	precuneus	0.642	Cognition.spatial	*n.s.*
14	R	temporooccipital junction	0.641	Cognition.spatial; emotion; action.observation; perception.vision shape; interoception. sexuality	Passive viewing; film viewing; affective pictures
15	R	superior frontal gyrus	0.640	*n.s.*	*n.s.*
16	L	inferior frontal gyrus	0.639	Cognition.language speech/semantics/syntas	Word generation (overt); reading (overt); semantic monitor/discrimination

17	R	supramarginal gyrus/temporo-occipital junction	0.639	Perception.audition; cognition.language speech	Film viewing; saccades; word generation (overt); pitch monitor/discrimination
18	R	precentral gyrus/inferior frontal gyrus, pars opercularis	0.638	*n.s.*	Stroop color; encoding; counting/calculation
19	R	lingual gyrus	0.634	*n.s.*	*n.s.*
20	L	calcarine cortex/precuneous	0.633	*n.s.*	*n.s.*
21	L	calcarine cortex/precuneous	0.632	*n.s.*	*n.s.*
22	L	parietal operculum cortex	0.630	Perception.audition; perception somesthesis.pain	Pain monitor/discrimination
23	R	precuneous	0.628	Cognition.social cognition	Theory of mind; face monitor/discrimination
24	L	fusiform cortex/lingual gyrus	0.628	Perception.vision	Drawing
25	L	calcarine cortex	0.628	*n.s.*	*n.s.*
26	R	precentral gyrus	0.627	Action.execution	Music production; Flexion/extension; finger tapping/button press
27	L	postcentral gyrus	0.624	Action.execution; perception somesthesis	Finger tapping/button press
28	R	caudate nucleus	0.620	Emotion.reward; cognition.reasoning	reward
29	R	PCC	0.619	Perception.vision	*n.s.*
30	L	precuneous	0.617	*n.s.*	*n.s.*
31	R	superior occipital cortex	0.617	Action.execution/motor learning/imagination; cognition.spatial	Drawing; imagined movement
32	R	superior occipital cortex	0.617	Cognition.social cognition	*n.s.*
33	L	(para)cingulate gyrus	0.617	Cognition.social cognition/reasoning; emotion.fear/reward; perception.gustation	Reward; taste; theory of mind
34	L	precuneous	0.617	Cognition.social cognition	*n.s.*
35	L	thalamus	0.616	Interoception; perception somesthesis.pain	*n.s.*
36	R	superior/middle temporal gyrus	0.612	Action.observation; perception.audition; cognition.language speech	passive listening; music production; face, pitch and tone monitor/discrimination;

					reading (overt); film viewing
37	L	superior occipital cortex	0.611	Action.execution/motor learning/imagination	Drawing; imagined movement
38	L	precuneous	0.609	Cognition.social cognition	Theory of mind; reasoning/problem solving
39	L	fusiform cortex	0.607	Cognition.language orthography/semantics; memory explicit	Naming (overt); encoding; semantic monitor/ discrimination
40	R	inferior temporal gyrus	0.597	Perception.audition	*n.s.*
41	R	superior occipital cortex	0.597	Cognition.social cognition	Theory of mind; Imagined objects/scenes
42	L	precuneous/PCC	0.592	Cognition.social cognition; emotion	Theory of mind
43	L	precuneous	0.592	Cognition.social cognition/language/memory explicit	Theory of mind; self-reflection
44	R	putamen	0.590	Emotion.reward; action.execution; perception somesthesis.pain	Reward; flexion/extension
45	L	parahippocampal gyrus/fusiform cortex	0.587	Cognition.language semantics and speech; cognition.memory explicit	Naming (overt); semantic monitor/discrimination
46	L	parahippocampal gyrus/PCC	0.586	Cognition. memory explicit	*n.s.*
47	R	superior occipital cortex	0.584	Cognition.reasoning and working memory	*n.s.*
48	L	TPJ	0.576	Cognition.social cognition	Theory of mind
49	R	supramarginal gyrus	0.575	Action.motor learning and execution; cognition.spatial	Drawing; anti-saccades; flanker; delayed match to sample; counting/calculation
50	R	paracingulate gyrus	0.574	Emotion.reward and fear; cognition.social cognition	reward
51	R	middle temporal gyrus	0.571	Perception.audition; cognition.social cognition	Passive listening; theory of mind
52	R	occipital fusiform gyrus	0.567	Action.observation	Passive viewing
53	L	precentral gyrus	0.558	*n.s.*	Flexion/extension

Abbreviations: vmPFC: ventral medial prefrontal cortex; TPJ: temporoparietal junction; PCC: posterior cingulate cortex; L: left; R: right; n.s: not significant.

Supplemental Figures

Figure S1. Illustration of the OPNMF factorization on the PANSS data

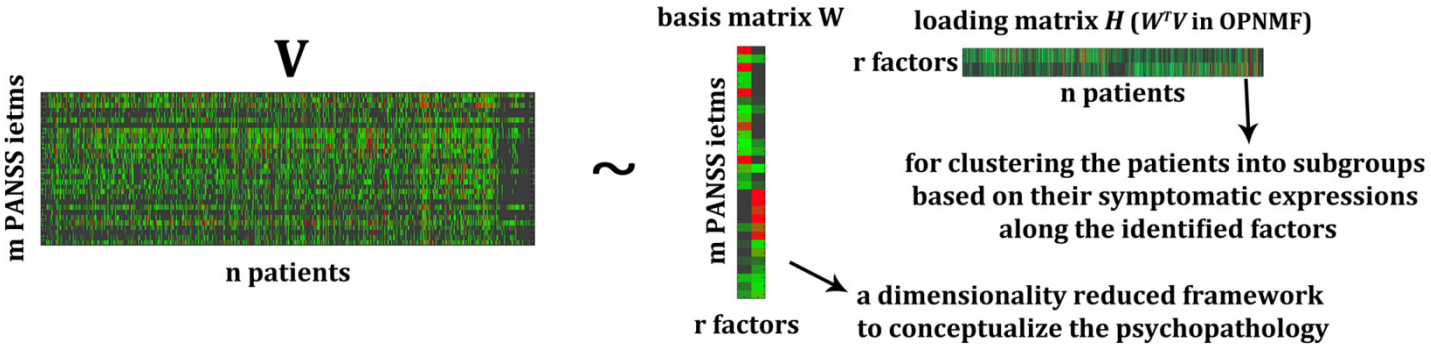


Figure S2. Flow chart for OPNMF evaluation procedures

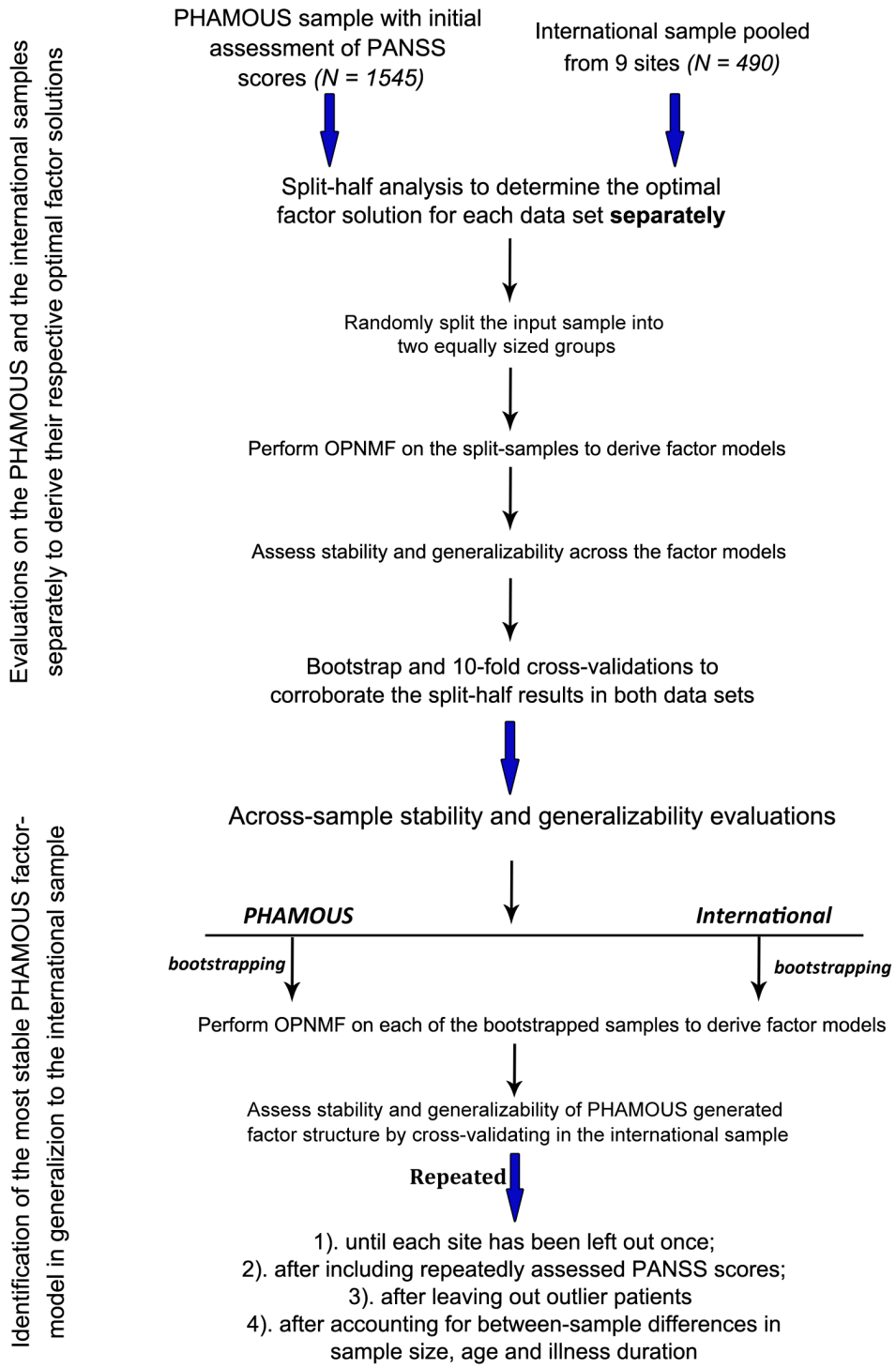
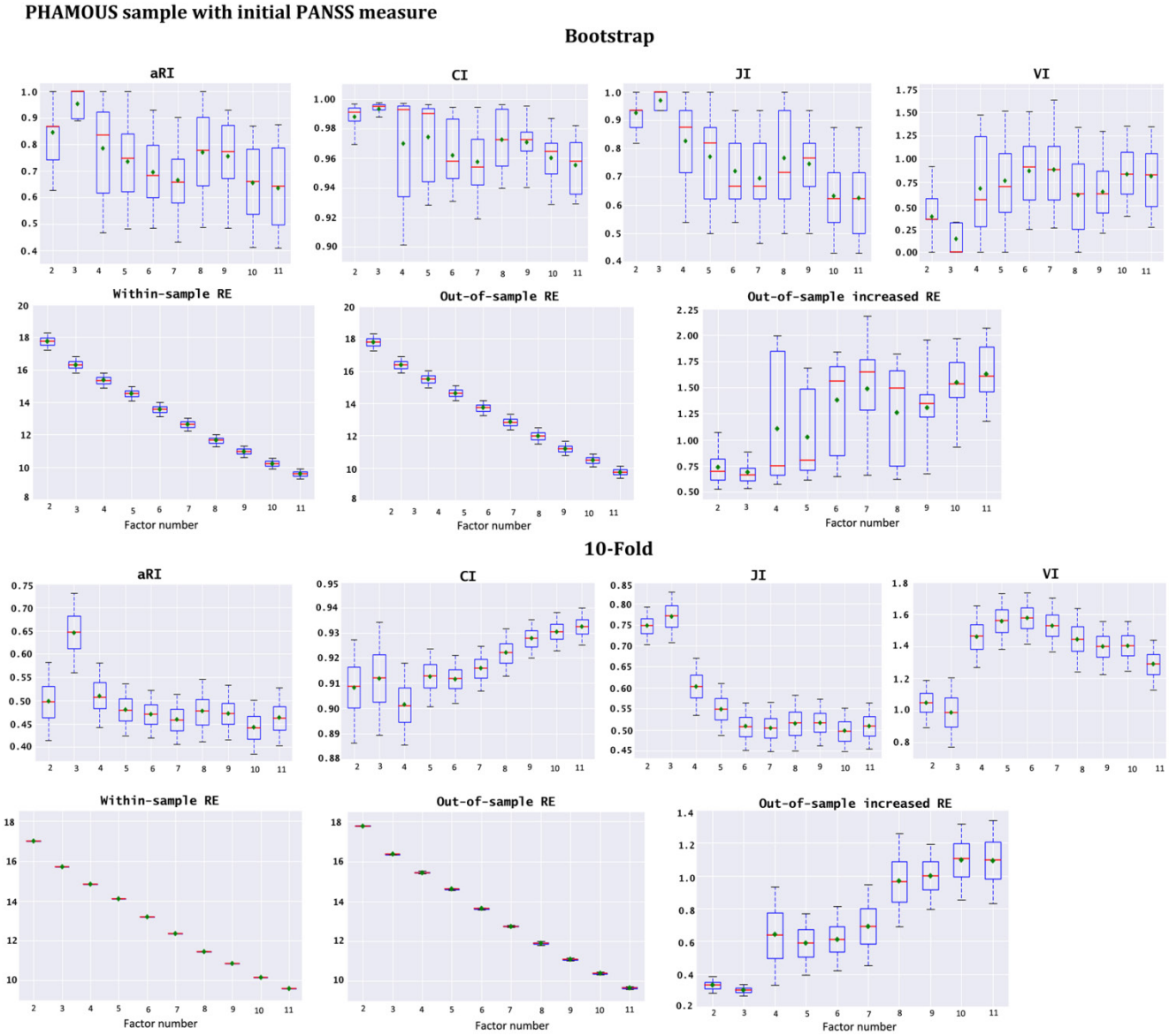
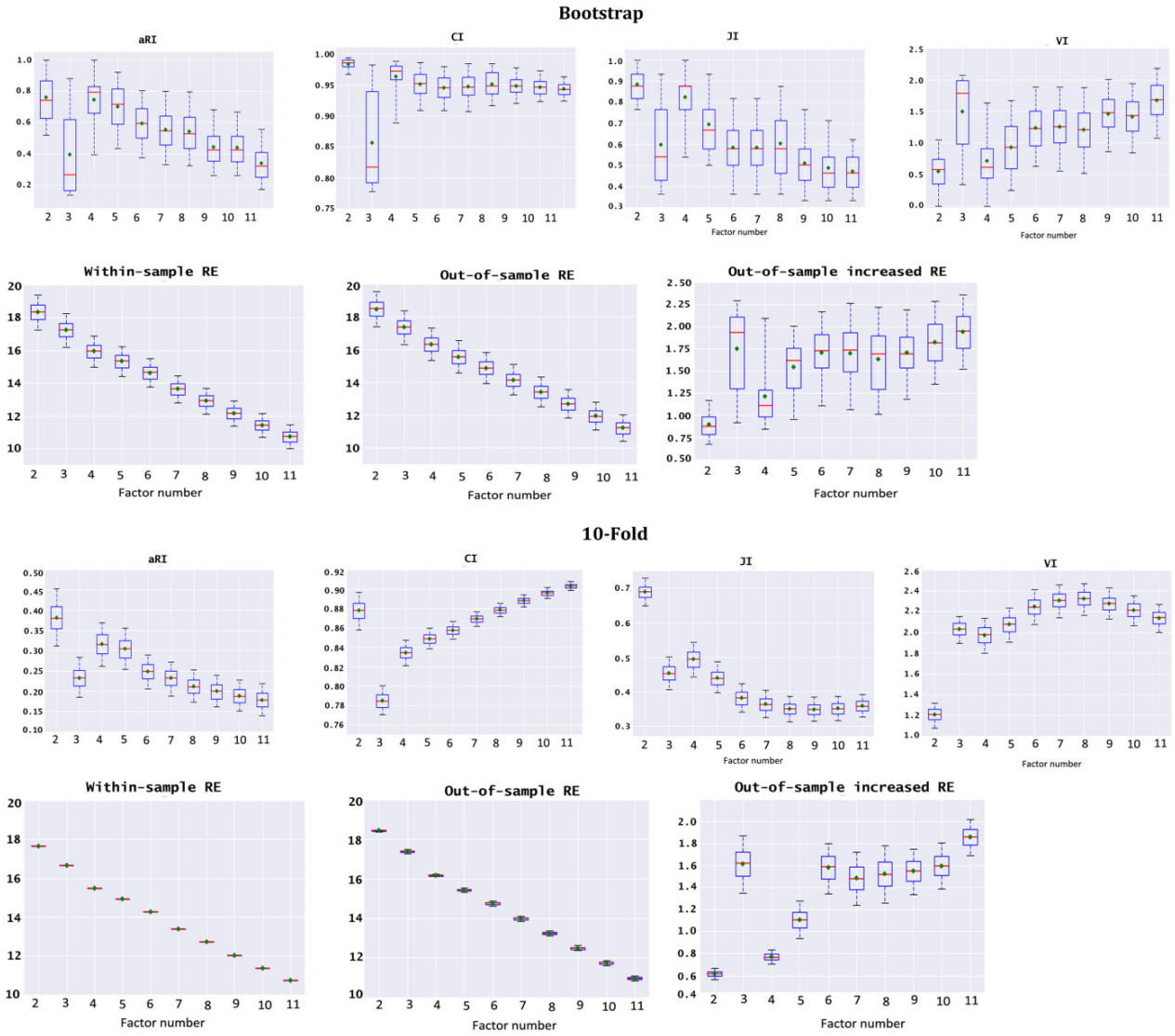


Figure S3. Bootstrap and 10-fold cross-validation (repeated for 10,000 times) of stability and generalizability for the factor-solutions derived by OPNMF in the PHAMOUS sample



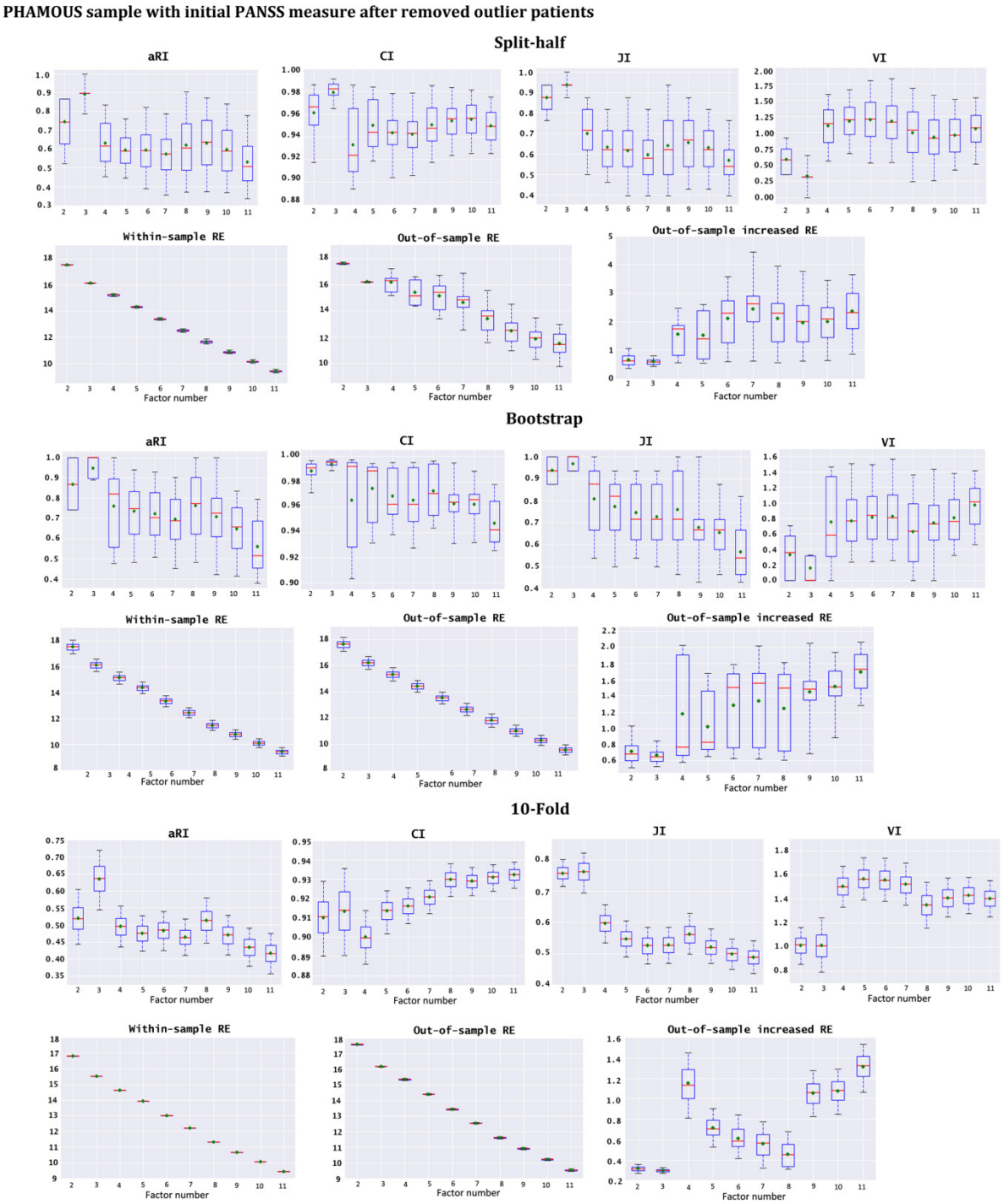
Box-plots show stability and generalizability results of the factor solutions from factor number 2 to 11. Higher values for adjusted RI and CI (upper row) indicate higher stability. Lower values for VI and out-of-sample increase in RE (bottom row) indicate better stability and generalizability, respectively. A 3-factor model was indicated as the best since both of the mean and median values for VI and out-of-sample increase in RE achieve the lowest, and the aRI and CI reach the highest at that point. For the box-plots, red line depicts the median, green diamond depicts the mean, whiskers represent the 5th and 95th percentiles. Abbreviations: aRI = adjusted Rand index, VI = variation of information, JI = Jaccard index, CI = concordance index, RE = reconstruction error.

Figure S4. Bootstrap and 10-fold cross-validations (repeated for 10,000 times) of stability and generalizability for the factor-solutions derived by OPNMF in the international sample



As shown, both of the two cross-validations inform that a 4-factor solution is the optimal. At point 4 the mean and median values of VI and out-of-sample increase in RE achieve the local minimum, while the aRI reaches the highest and the CI reaches the local maximum in bootstrap analysis. For 10-fold cross-validation, except for CI which does not point to a specific factor number, the other three indices all arrive at an optimal solution of factor number 4 (the mean and median values of VI and out-of-sample increase in RE achieve the local minimum, while the aRI reaches the local maximum). For the box-plots, red line depicts the median, green diamond depicts the mean, whiskers represent the 5th and 95th percentiles. Abbreviations: aRI = adjusted Rand index, VI = variation of information, JI = Jaccard index, CI = concordance index, RE = reconstruction error.

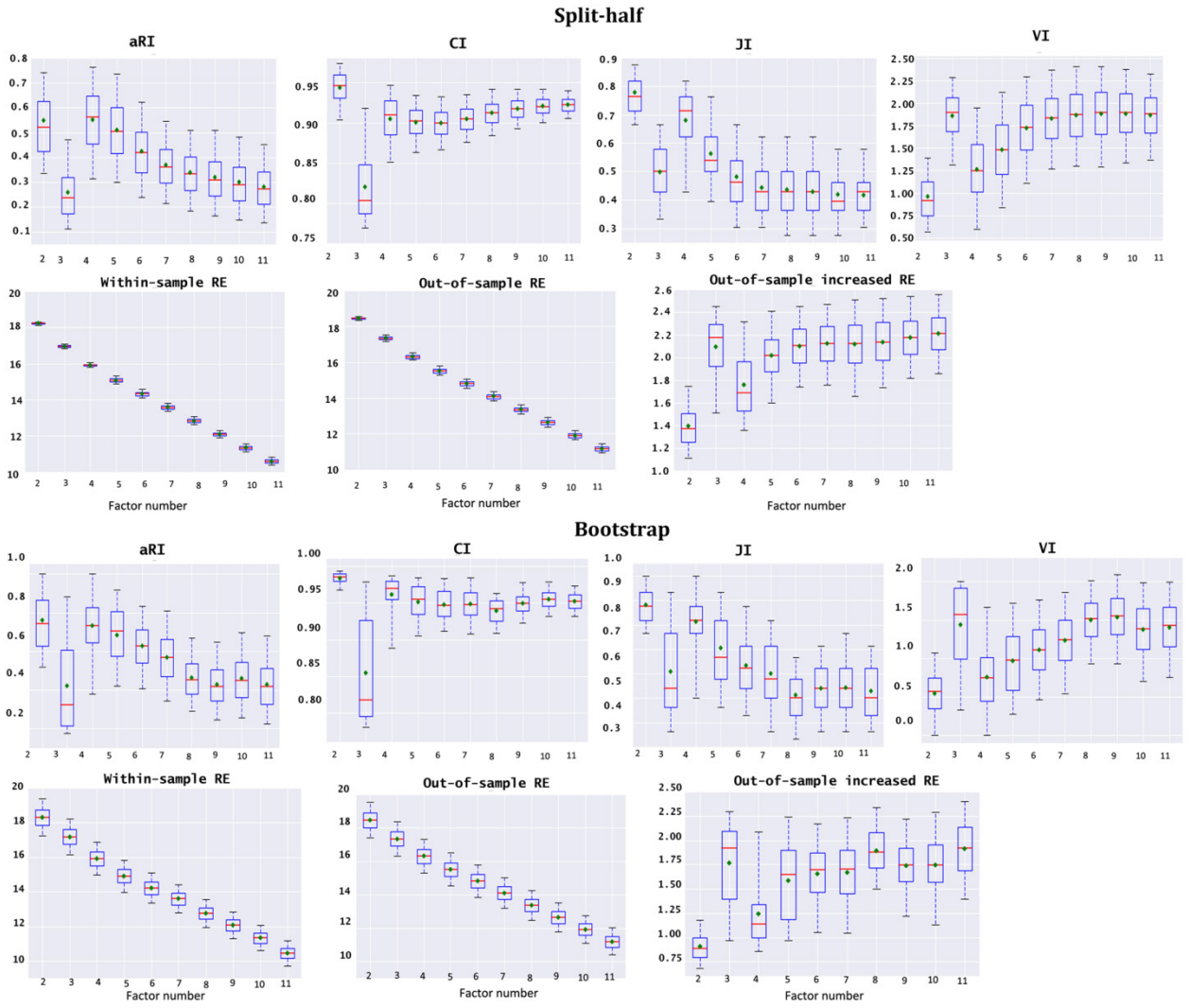
Figure S5. Split-half, bootstrap, and 10-fold cross-validations (repeated for 10,000 times) of stability and generalizability for the factor-solutions derived by OPNMF in the PHAMOUS sample after removed outlier patients using themedian absolute deviation method



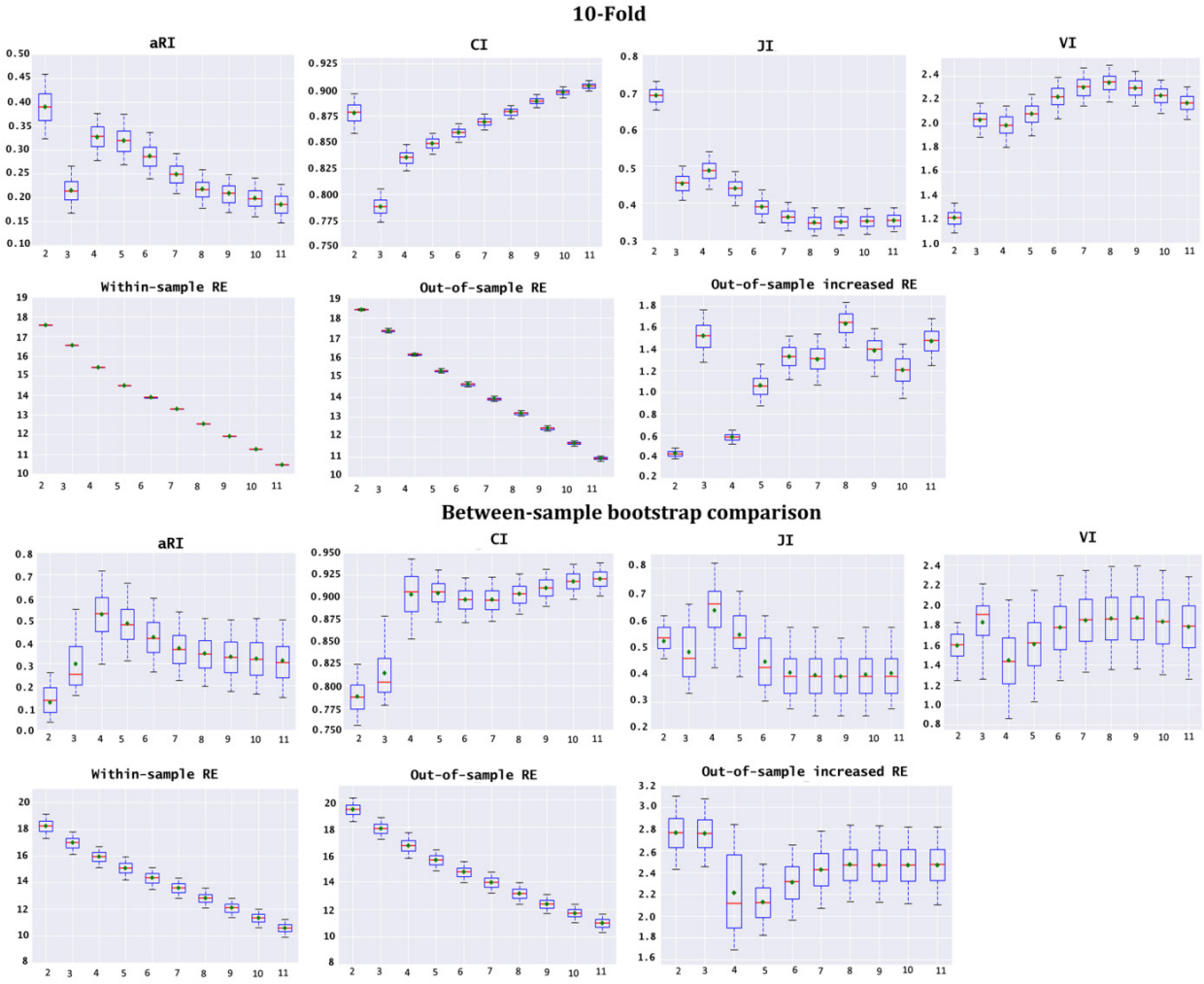
As shown, a 3-factor model was indicated as the best in all of the three manners of cross-validation as both of the mean and median values for VI and out-of-sample increase in RE achieve the lowest, and the aRI and CI (for 10-fold, it reaches the local maximum) reach the highest at that point. For the box-plots, red line depicts the median, green diamond depicts the mean, whiskers represent the 5th and 95th percentiles. Abbreviations: aRI = adjusted Rand index, VI = variation of information, JI = Jaccard index, CI = concordance index, RE = reconstruction error.

Figure S6A-S6B. Split-half, bootstrap (S6A), and 10-fold cross-validations (S6B) in the international sample, as well as the between-sample bootstrap comparison (repeated for 10,000 times; S5B) of stability and generalizability for the factor-solutions derived by OPNMF after removed outlier patients using the median absolute deviation method

S6A.



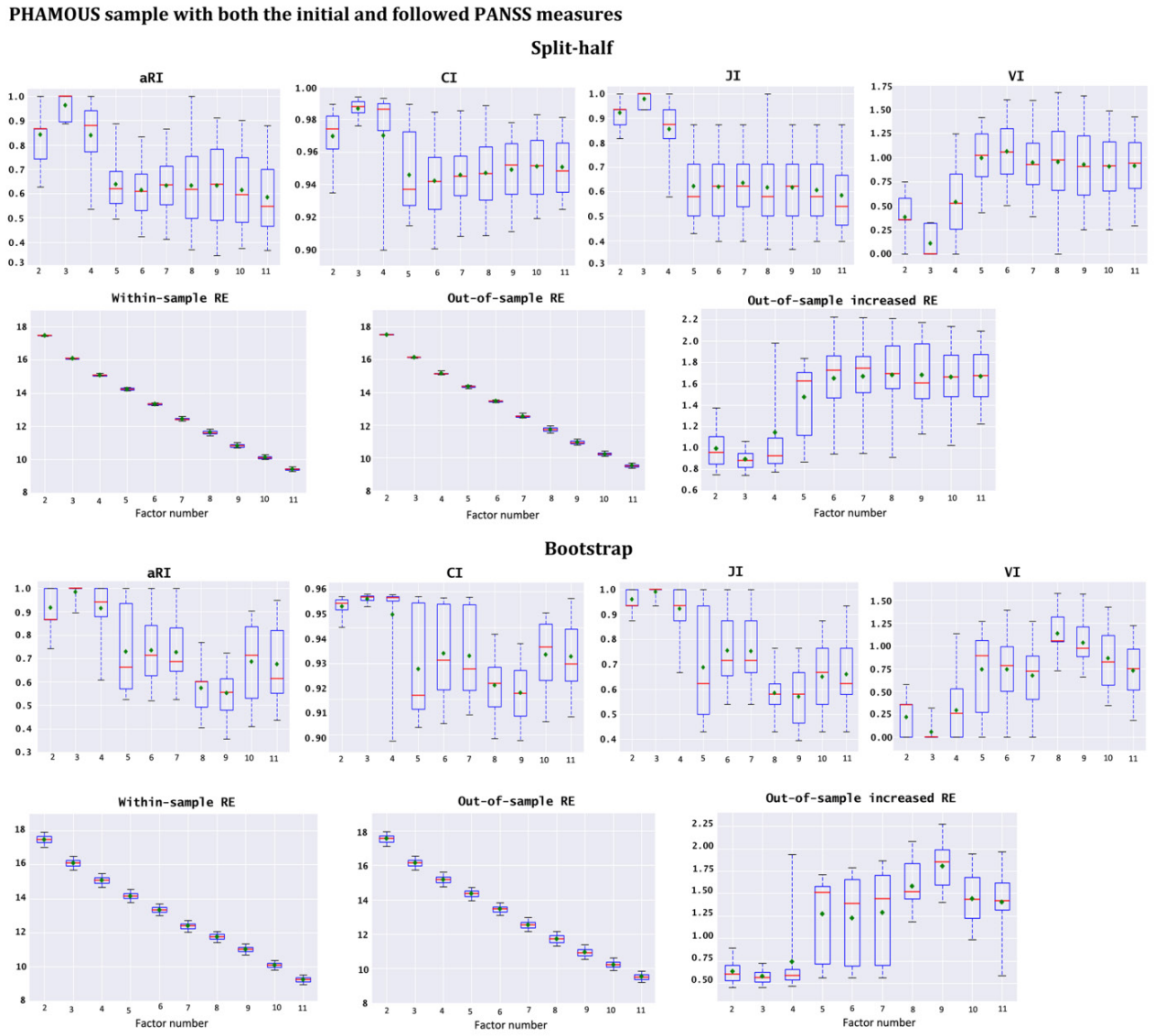
S6B.



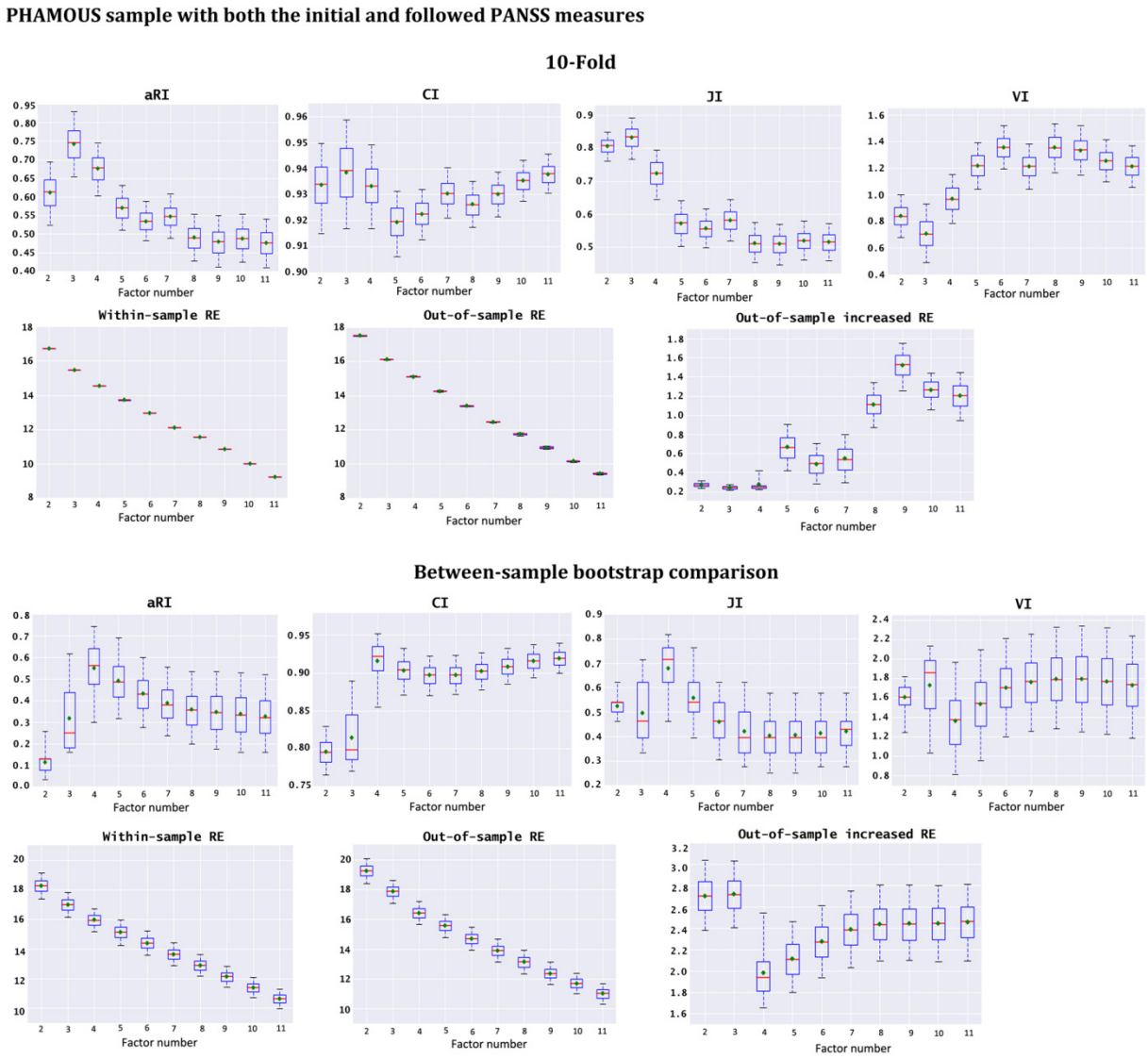
As shown, 4 factor solution is optimal for the international sample as well as generalizing cross the two datasets because at that point the median values of VI and out-of-sample increase in RE achieve the local minimum, while the aRI and the CI (except for that in 10-fold cross-validation) reach the local maximum in all of the three (split-half, bootstrap and 10-fold) cross-validation strategies in the international sample, as well as in the between-sample (PHAMOUS vs. international) bootstrap comparison analysis. For the box-plots, red line depicts the median, green diamond depicts the mean, whiskers represent the 5th and 95th percentiles. Abbreviations: aRI = adjusted Rand index, VI = variation of information, JI = Jaccard index, CI = concordance index, RE = reconstruction error.

Figure S7A-S7B. Split-half, bootstrap (S7A), and 10-fold cross-validations (repeated for 10,000 times; S7B) in the PHAMOUS sample with both the initial and follow-up PANSS measures, as well as the between-sample bootstrap comparison (S7B) of stability and generalizability for the factor-solutions derived by OPNMF

S7A.

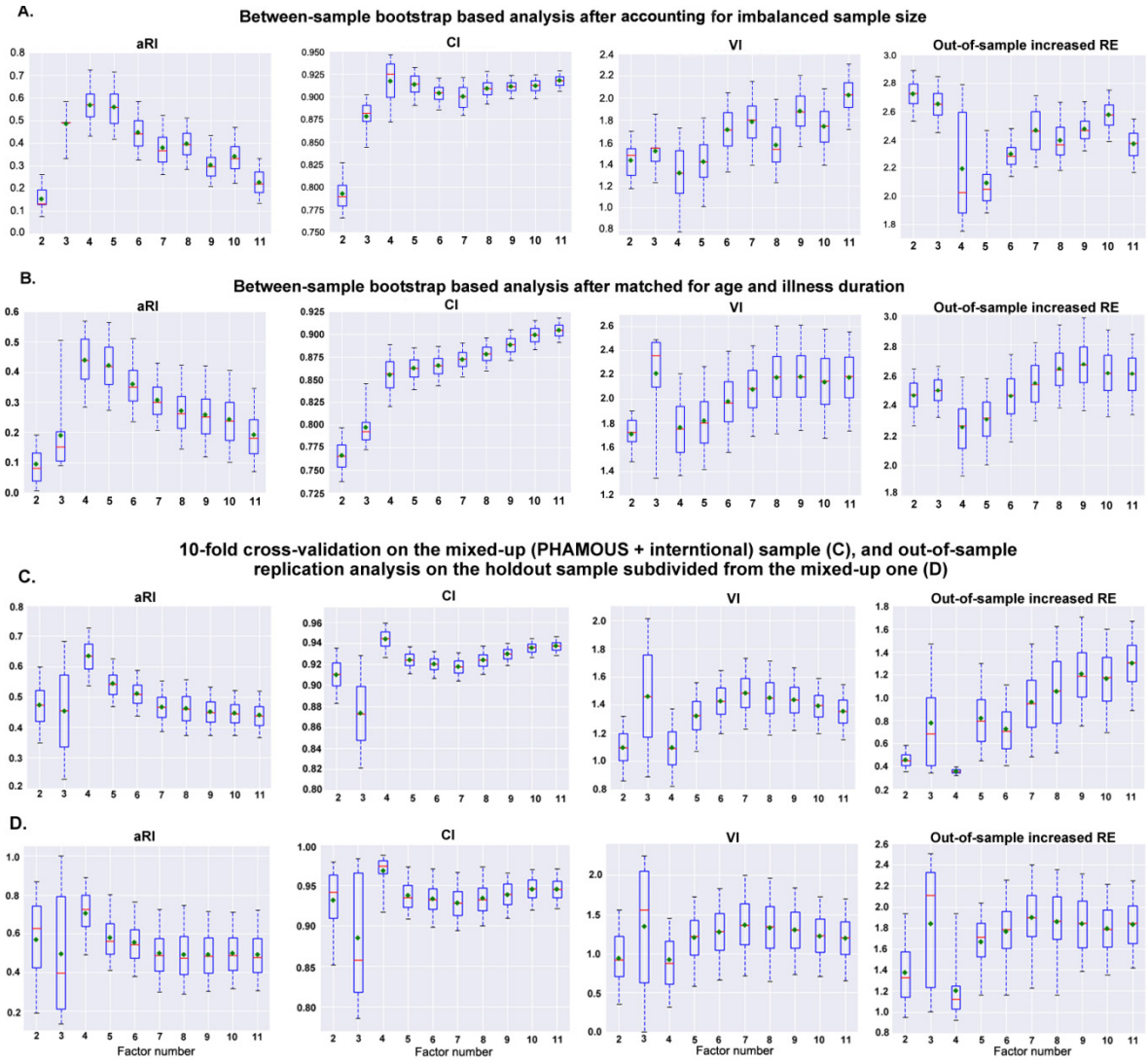


S7B.



As shown, factor number 3 is the optimal solution for the PHAMOUS sample in split-half, bootstrap and 10-fold cross-validations because at that point the mean and median values of VI and out-of-sample increase in RE achieve the lowest, while the aRI and CI reach the highest. In between-sample comparison analysis, a 4 factor-structure is also shown as the best when generalizes from the large, homogeneous PHAMOUS sample to the international sample since both of the mean and median values for VI and out-of-sample increase in RE achieve the lowest, and the aRI and CI reach the highest at factor number 4.For the box-plots, red line depicts the median, green diamond depicts the mean, whiskers represent the 5th and 95th percentiles. Abbreviations: aRI = adjusted Rand index, VI = variation of information, JI = Jaccard index, CI = concordance index, RE = reconstruction error.

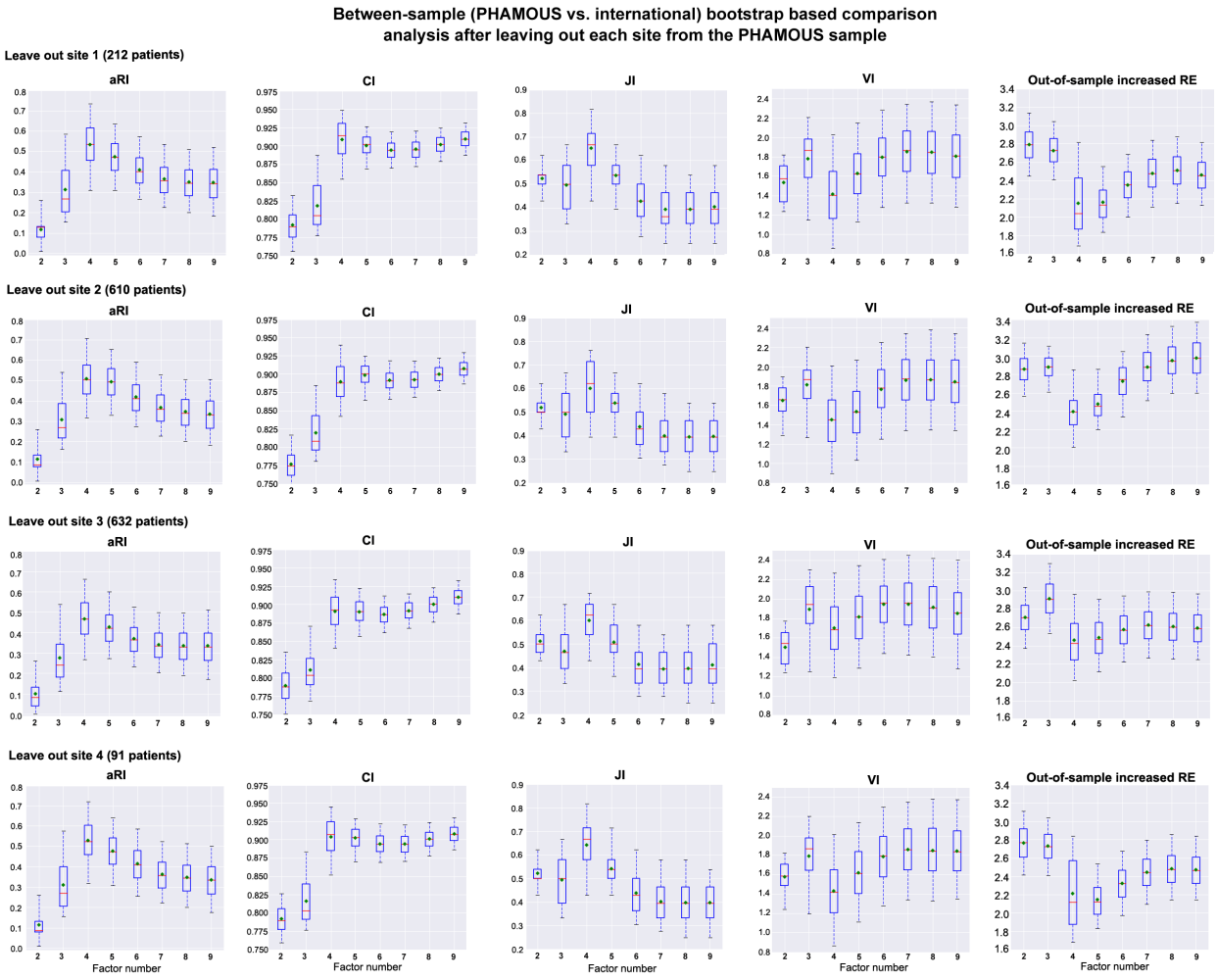
Figure S8. Between-sample bootstrap analysis after accounting for age and illness duration differences, as well as sample size difference between the two datasets



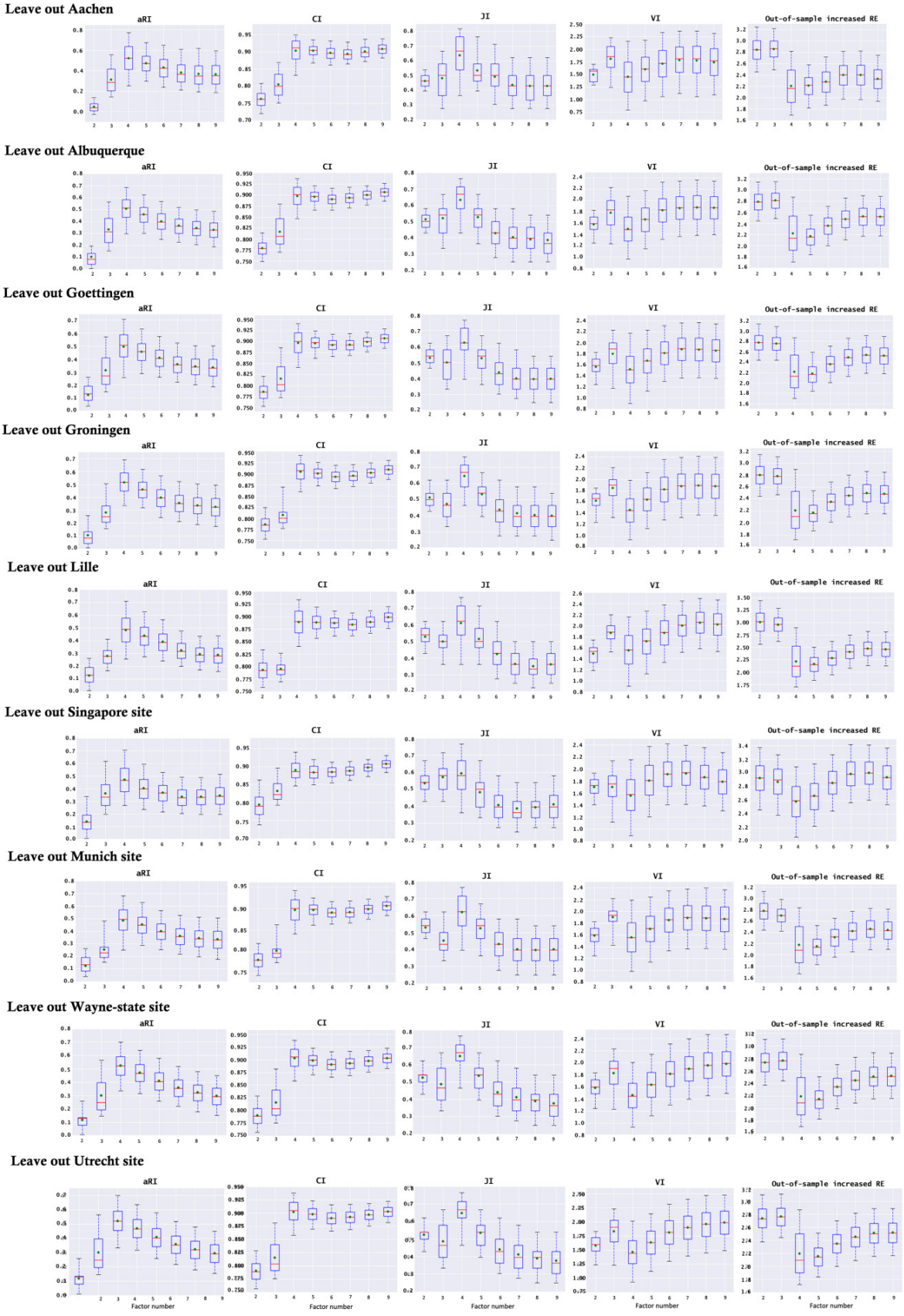
As shown, a four-factor model stays as the optimal solution in generalizing to the international data set after the differences in sample size, age and illness duration between the PHAMOUS and the international sample were accounted. Both of the mean and median values for variation of information and out-of-sample increase in reconstruction error achieve the lowest, and the adjusted Rand index and concordance index reach the highest at factor number 4. Red line depicts the median, green diamond depicts the mean, whiskers represent the 5th and 95th percentiles. Abbreviations: aRI = adjusted Rand index, VI = variation of information, JI = Jaccard index, CI = concordance index, RE = reconstruction error.

Figure S9A-S9B. Between-sample (PHAMOUS vs. international) bootstrap based comparison analysis after leaving out each site from the PHAMOUS sample (S9A), or the international dataset (S9B) consecutively to confirm that the best generalizable factor-model is independent of geographical regions

S9A



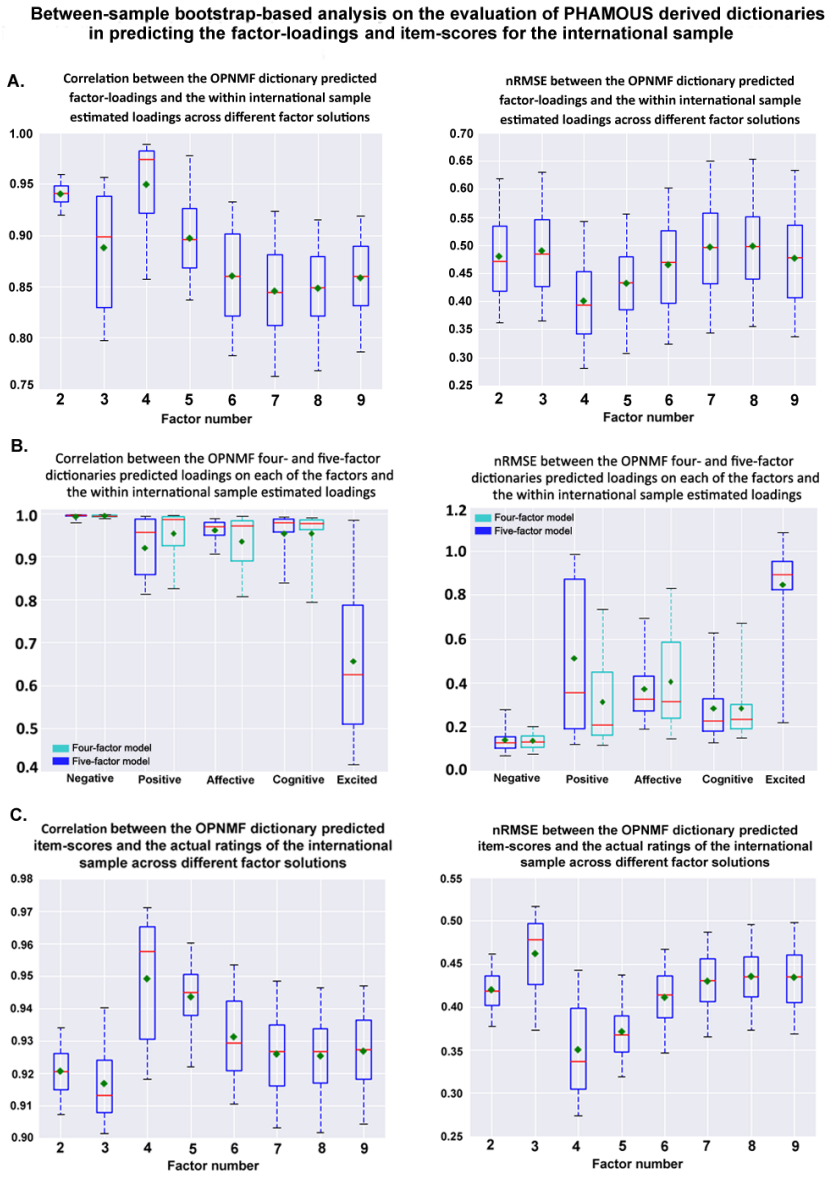
S9B



All of the 13 leave-one site-out analyses indicate a 4 factor-model is the most generalizable solution across the two datasets. Seen from the figure, the mean and median values of VI and the median value of out-of-sample increase in RE achieve the local minimum, while the aRI reaches the highest and the CI reaches the local maximum no matter which site

was left out from the PHAMOUS or the international dataset. For the box-plots, red line depicts the median, green diamond depicts the mean, whiskers represent the 5th and 95th percentiles. Abbreviations: aRI = adjusted Rand index, VI = variation of information, JI = Jaccard index, CI = concordance index, RE = reconstruction error.

Figure S10. Between-sample (PHAMOUS vs. international) bootstrap based analysis for assessing the stability and accuracy of PHAMOUS generated dictionaries in predicting the loadings and item-scores for the international sample



Note: The newly emerged fifth factor in the OPNMF five-factor model represents excited symptoms primarily including poor impulse control and excitement items.

Figure S11. Between-sample (PHAMOUS vs. international) bootstrap based analysis for assessing the stability and accuracy of PHAMOUS generated dictionaries in predicting the loadings and item-scores for each individual site in the international sample

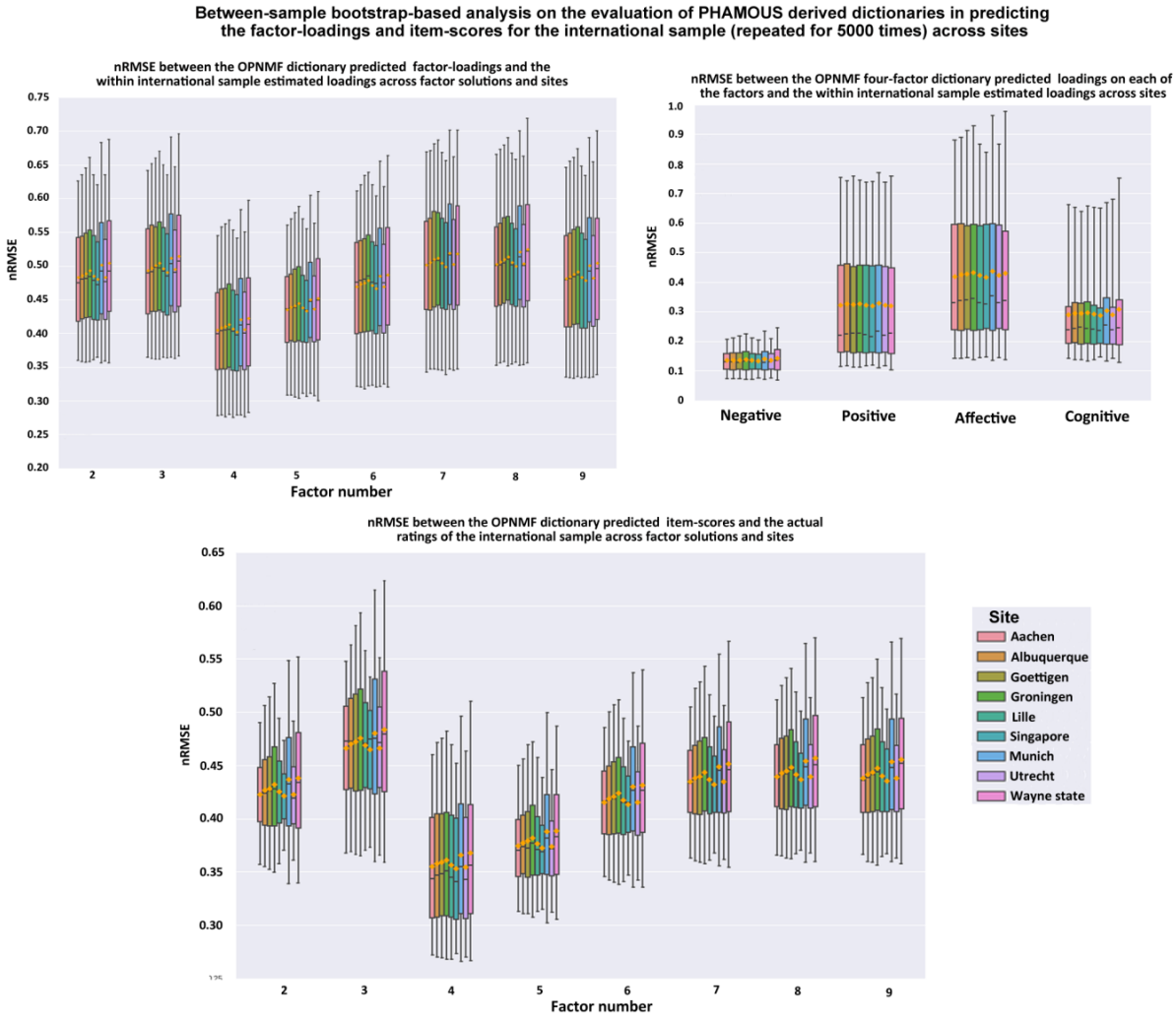
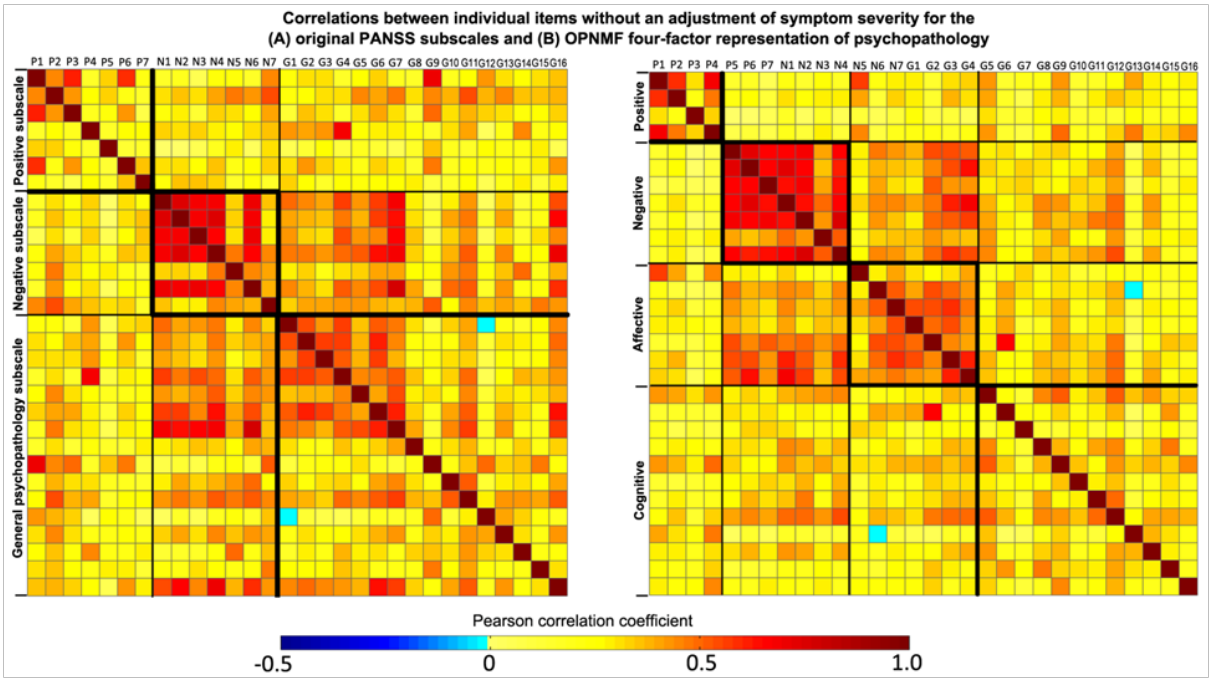
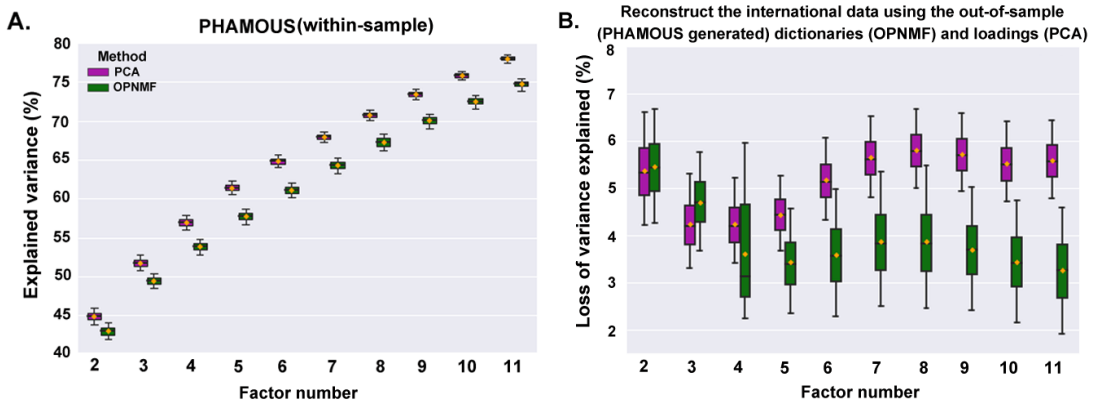


Figure S12. Inter-item correlations for the original PANSS subscales and OPNMF representation of psychopathology without adjusting for general symptom severity (total PANSS score)



Correlation strength is color-coded (light yellow to red: positive correlations; cyan to blue: negative correlations).

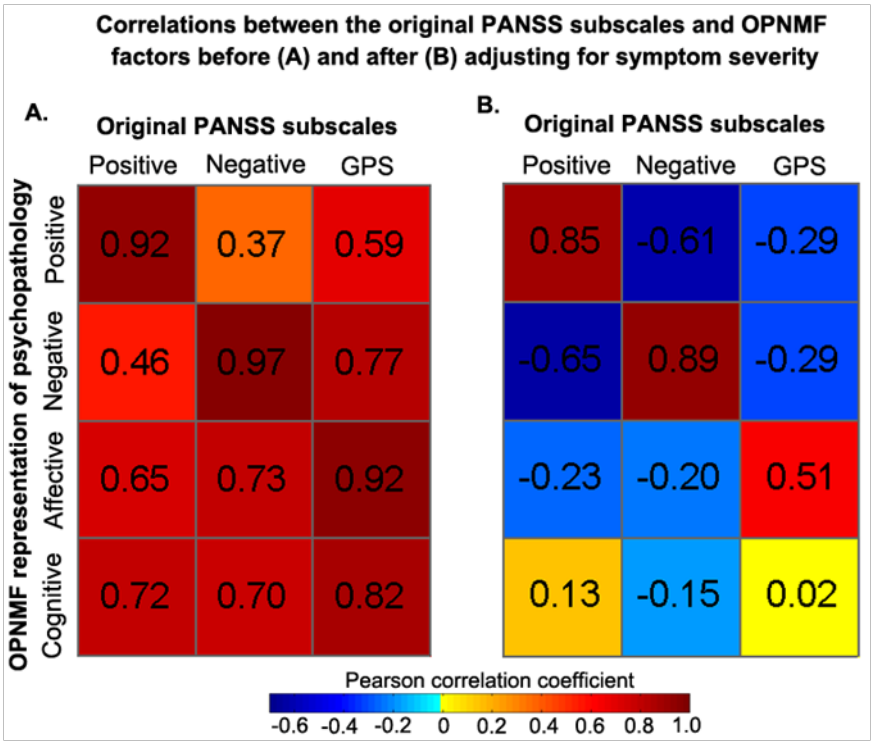
Figure S13. Quantitative comparison of PCA and OPNMF derived factor models based on explained variance



A) Within-sample explained variance (EV) for the matrix reconstructed by OPNMF dictionary and the PCA loadings.

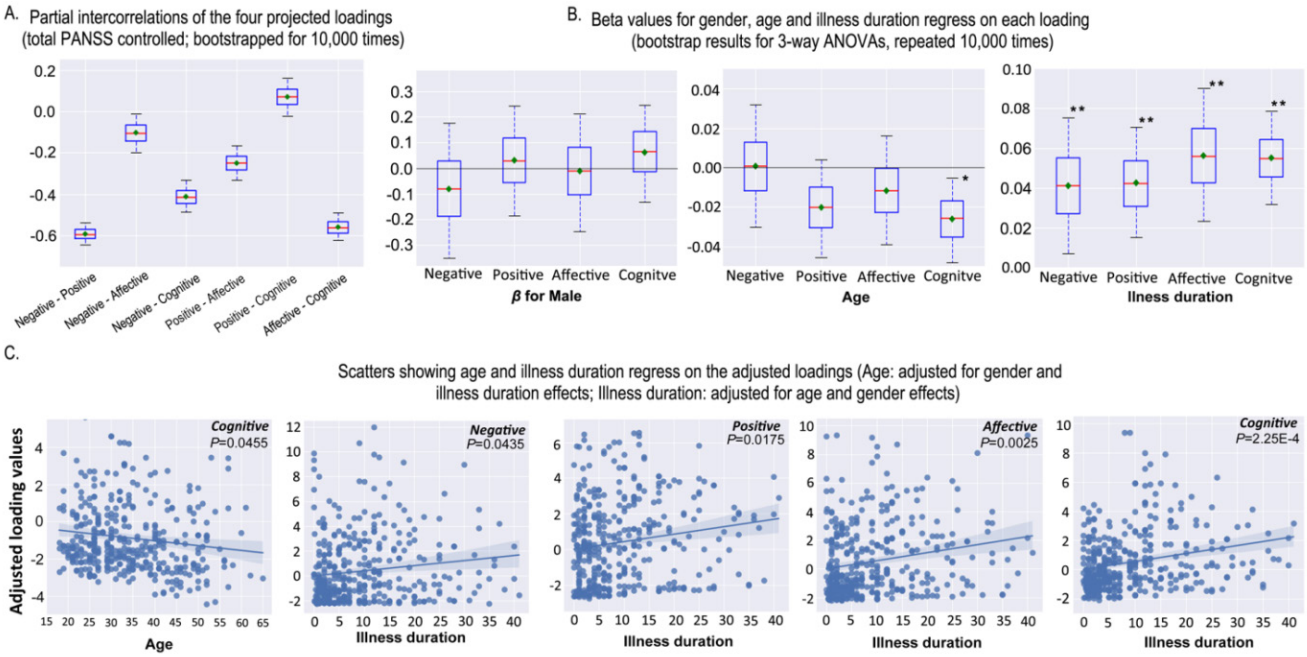
B) A higher loss of EV indicates worse generalizability. OPNMF is with better generalization performance with lower loss of EV, especially for the four factor model which achieved the local minimum.

Figure S14. Correlations between the three original PANSS subscales and OPNMF representation of psychopathology with and without controlling for general symptom severity (total PANSS score)



Correlation strength is color-coded (light yellow to red: positive correlations; cyan to blue: negative correlations).

Figure S15. Relationship between factors, demographic and clinical information without controlling for symptom severity



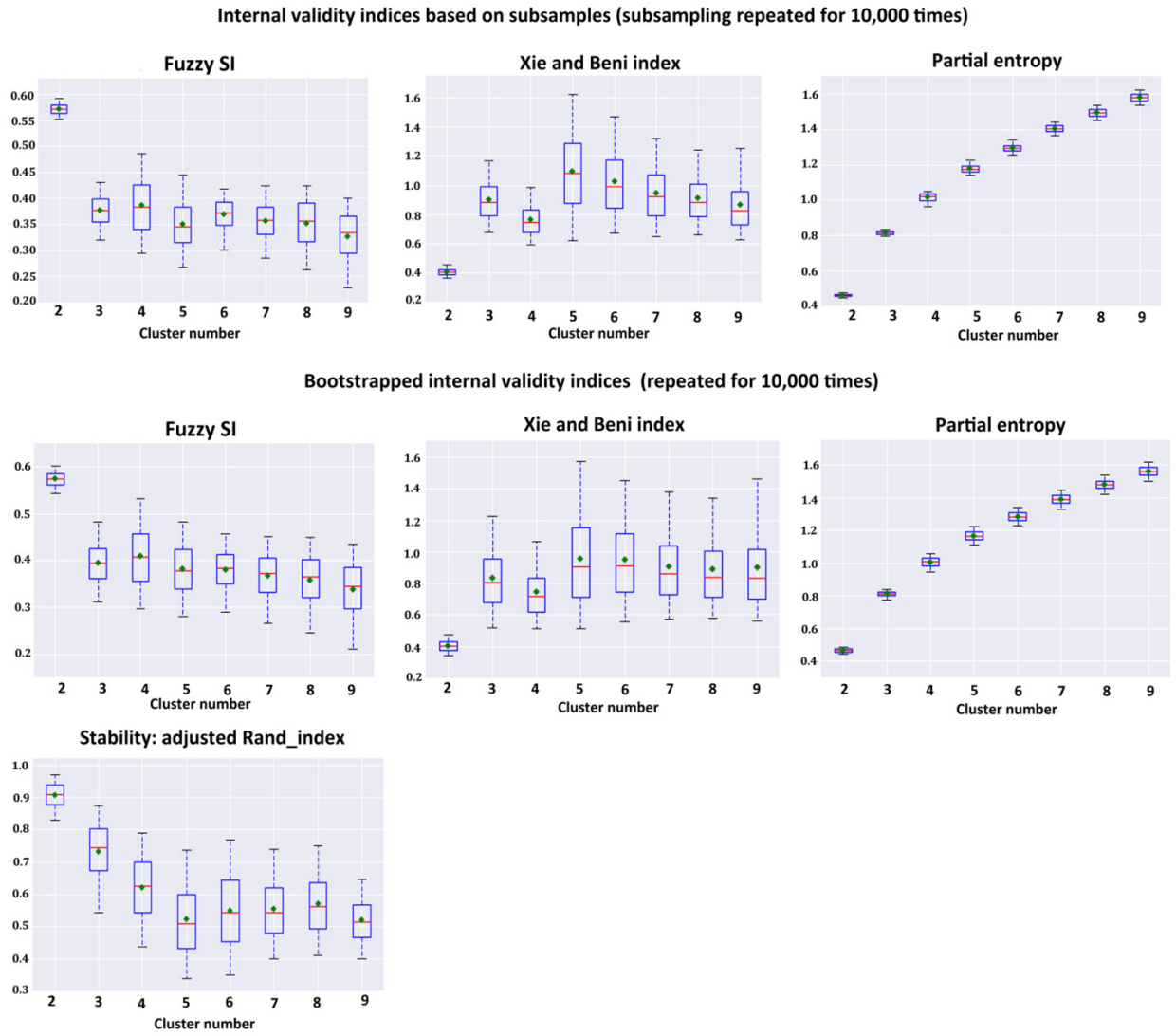
(A) Partial intercorrelation of the four factor-loadings after controlling for overall symptom severity (total PANSS score). Box-plot shows the bootstrap results (10,000 replication times) for the partial intercorrelation between the 4-factor loadings. Red line depicts the median, green diamond depicts the mean, whiskers represent the 5th and 95th percentiles.

B-C. Effects of socio-demographic and clinical features on the 4-factor loadings

(B) Bootstrap results for the 3-way ANOVA analysis. Samples were drawn from the original loading matrix for 10,000 times and then the ANOVAs were repeated on these bootstrapped samples. Boxes refer to the beta values. Red line depicts the median, green diamonds the mean. *(*Mdn*, $p < .05$), **(*M* and *Mdn*, $p < .05$).

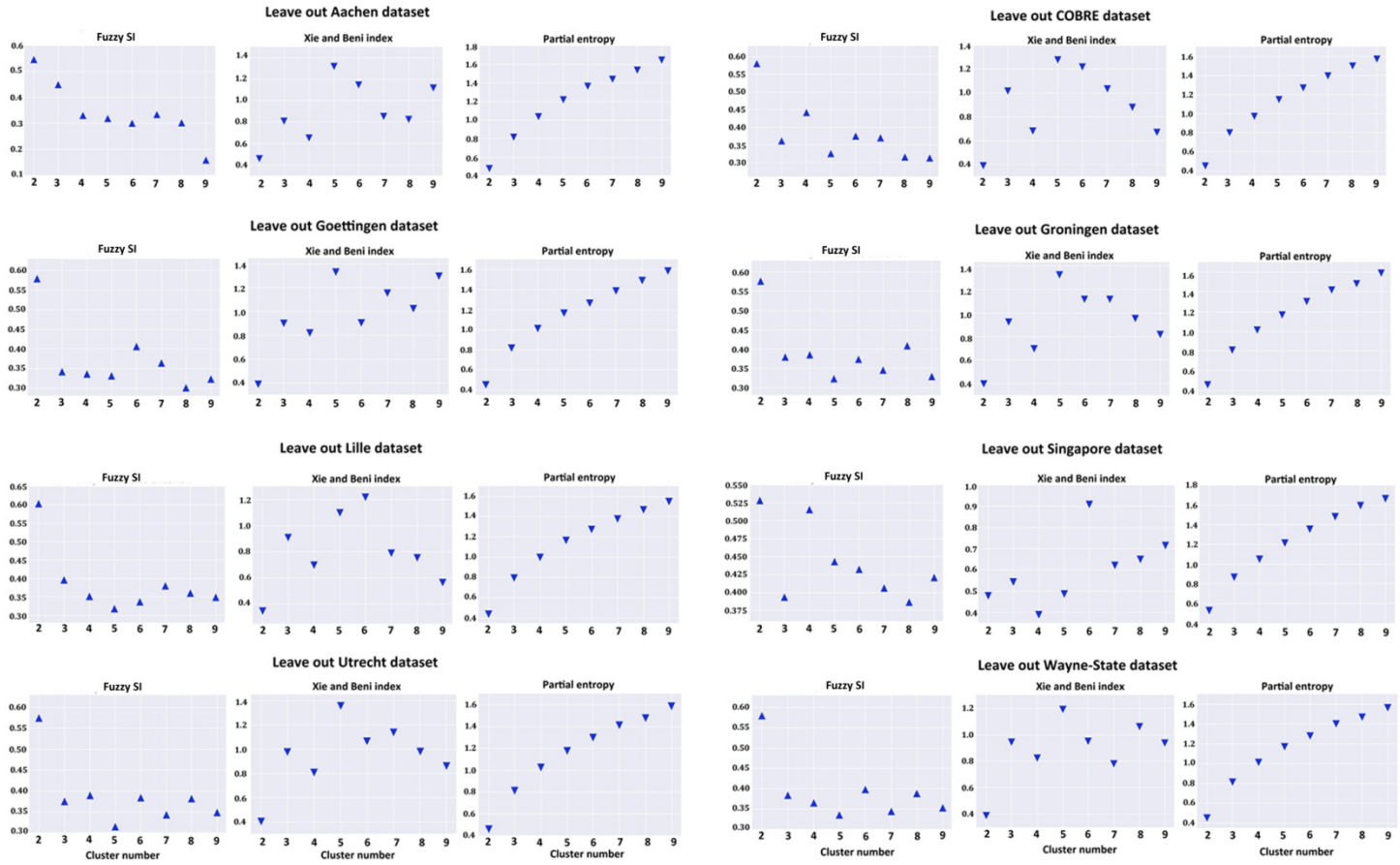
(C) Results of the 3-way ANOVA analysis based on the original loading matrix. Scatters show significant negative association between age (adjusted for gender and illness duration) and the cognitive loading, and significant positive associations between illness duration and the four factor-loadings after adjusted for age and gender. Regression lines were depicted with 95% confidence Interval on the fitted values.

Figure S16. Assessment of clustering stability based on subsampling and bootstrap resampling techniques. Internal validity indices were calculated for both of the subsamples and bootstrapped samples



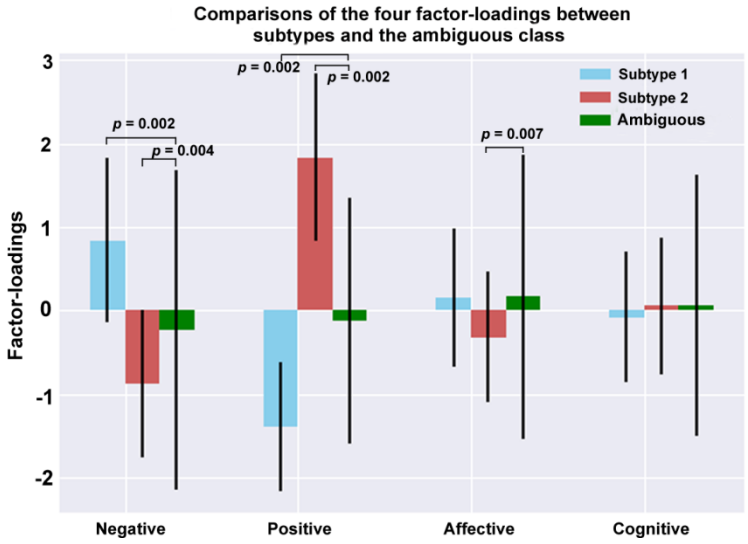
The boxplots show the values of the validity indices and adjusted Rand index (aRI) from cluster number 2 to 9. Higher values of fuzzy SI and lower values of XB and PE indicate a better clustering quality, while a higher aRI refers to better stability of the cluster partitions The mean and median maximum for fuzzy SI and the minimums for XB and PE were all on the point of 2, suggesting that a 2 cluster solution is the best in representing the subgroups within the schizophrenia patient population regardless of the original data that are perturbed. The cluster number 2 also reaches the highest aRI in bootstrap analysis among the tested cluster numbers. aRI reflects the convergent assignment of the patient-pairs to the clusters between the bootstrapped samples and the original sample (whole patients). Abbreviations: SI = Silhouette index, XB = Xie and Beni index, PE = partition entropy.

Figure S17. Leave-one-site-out validation for the optimal cluster solution by calculating the three internal validity indices upon the removal of each site from the international dataset consecutively



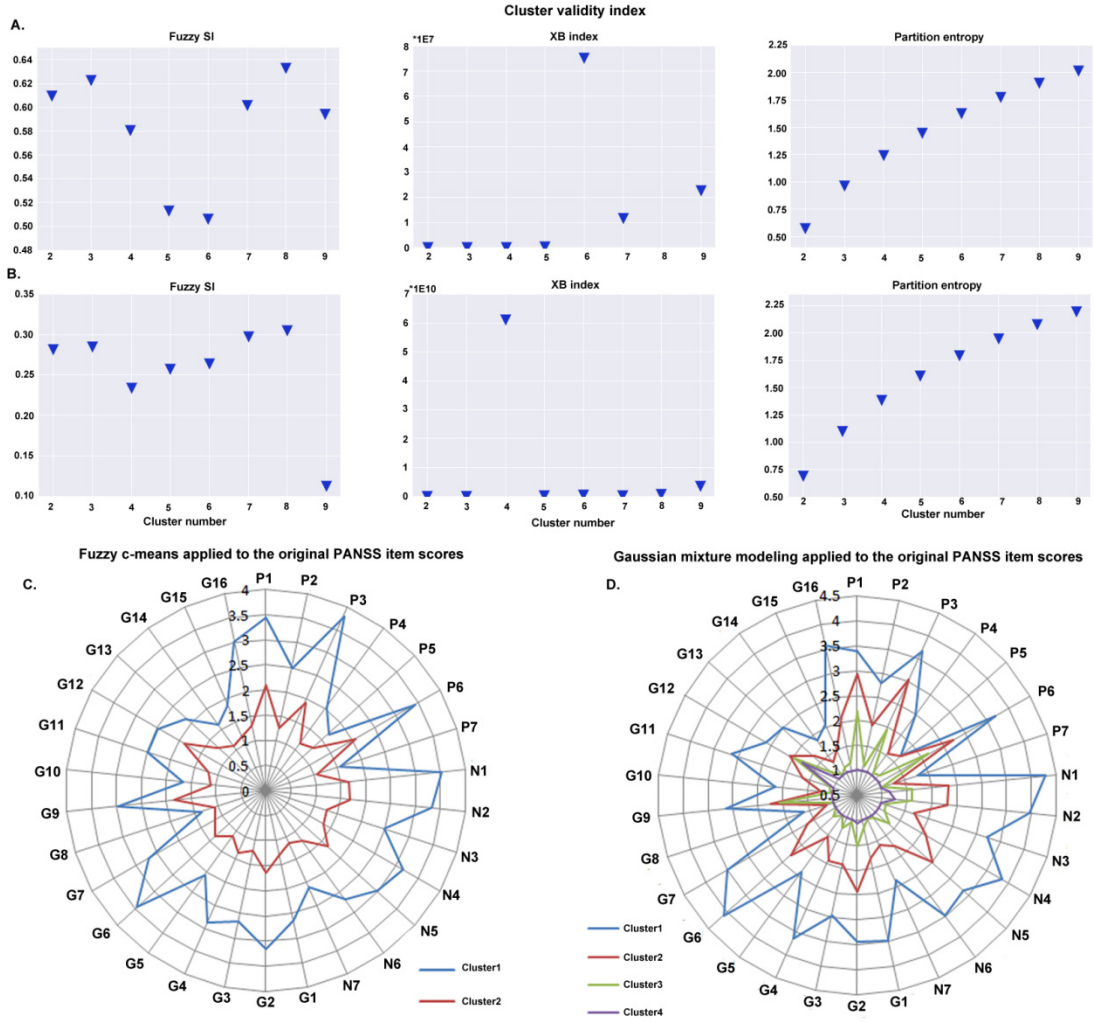
The plots show the values of the validity indices from cluster number 2 to 9. Higher values of fuzzy SI (in triangle) and lower values of XB and PE (in inverted triangle) indicate a better clustering quality. The maximums for FSI and the minimums for XB and PE were all on the point of 2, suggesting that a 2 cluster solution is the best in representing the subgroups within the schizophrenia patient population no matter which site is excluded from the international sample. Abbreviations: SI = Silhouette index, XB = Xie and Beni index, PE = partition entropy.

Figure S18. Comparisons of the four adjusted factor-loadings between the core subtypes and the ambiguous patient class



Factor-loadings (adjusted) that significantly differed between the subtypes and the class of 25% ambiguous cases were marked with p values provided (10,000 permutation tests with shuffled cluster labels). Error-bar: standard deviation

Figure S19. Clustering the original and the residuals of the 30 individual PANSS item-scores



A) Fuzzy c-means yielded an optimal two-cluster solution (XB index and partition entropy) for clustering the original 30 PANSS item-scores

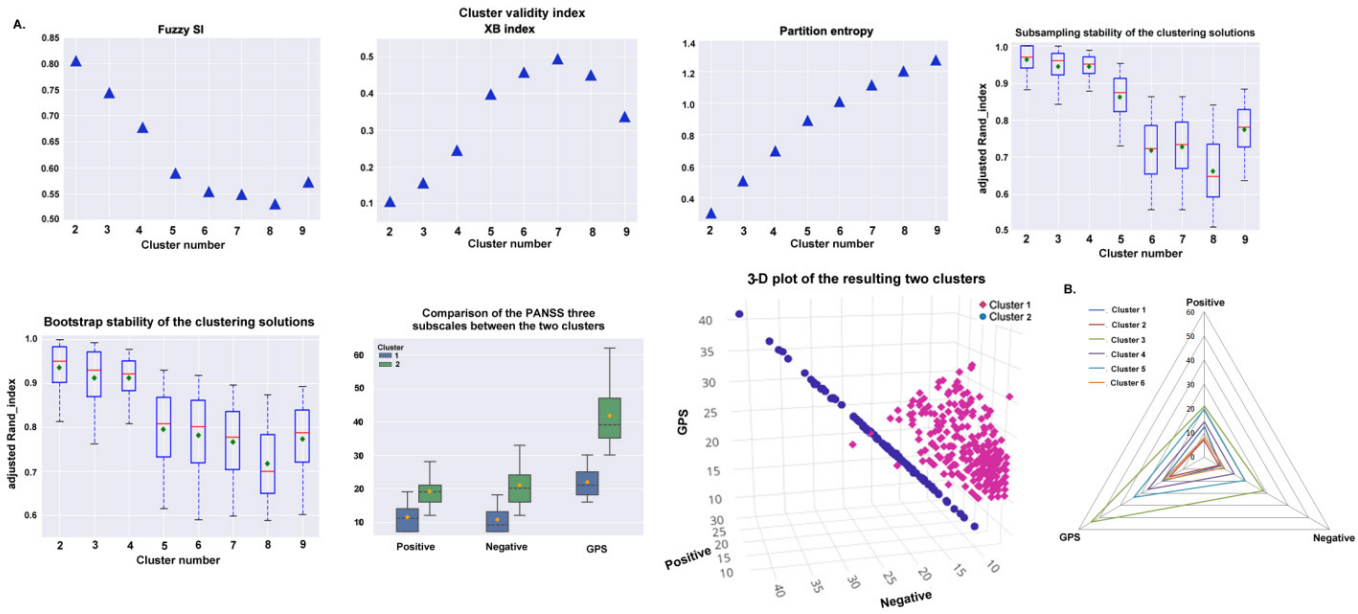
B) Although XB index and partition entropy pointed to a two-cluster solution, values of the internal validity index of XB are extremely large indicating poor clustering quality and

C) The clusters were mainly driven by overall symptom severity such that scores for the all 30 items were either high or low, respectively

D) The four clusters identified by GMM were reflecting to overall symptom severity

Higher values of fuzzy SI (in triangle) and lower values of XB and PE (in inverted triangle) indicate a better clustering quality. Abbreviations: SI = Silhouette index, XB = Xie and Beni index, PE = partition entropy.

Figure S20. Clustering the original three PANSS subscales without adjusting for age, gender, illness duration and total PANSS score

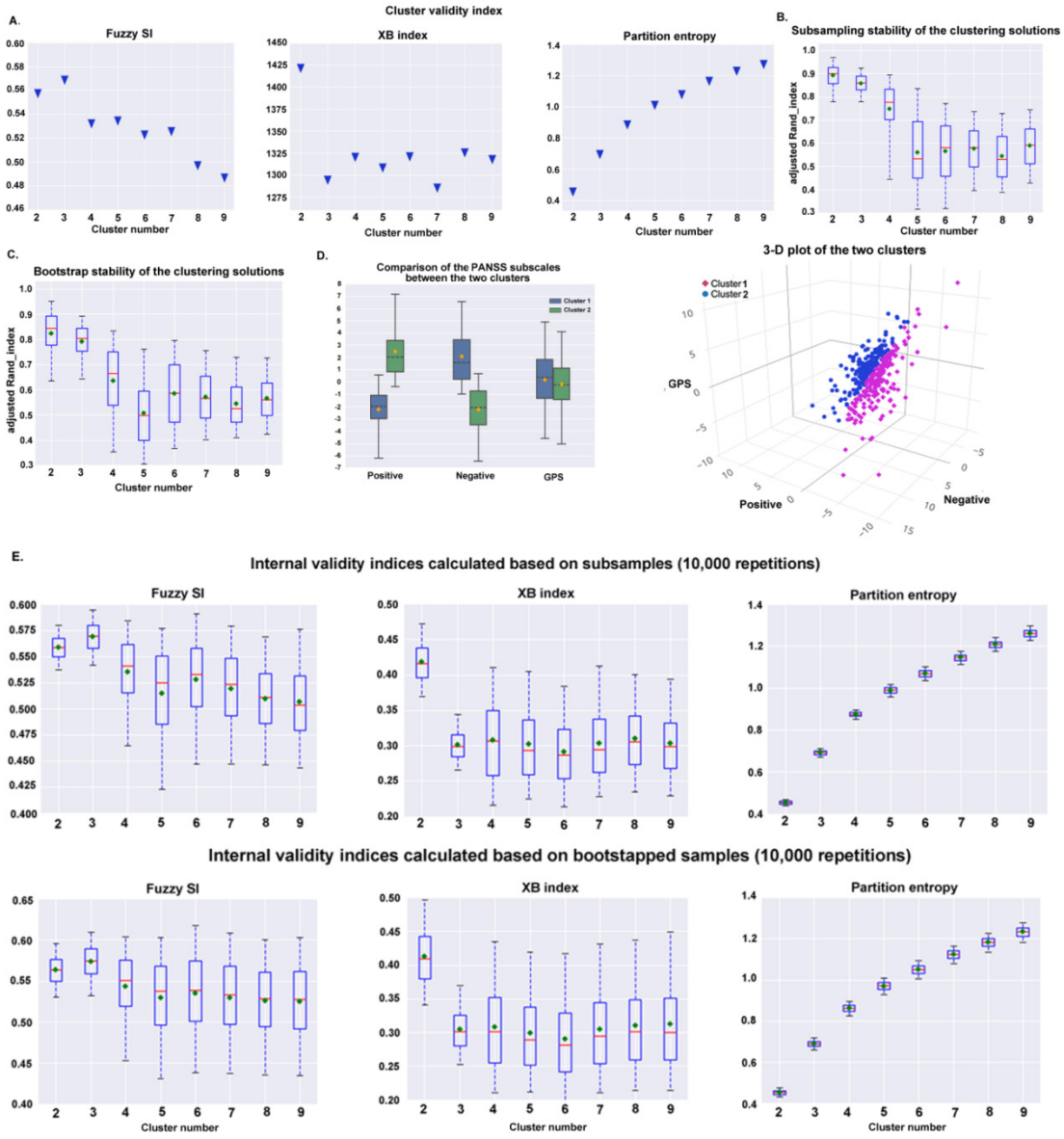


A) Fuzzy c-means results: Higher values of fuzzy Silhouette index (SI), lower values of Xie and Beni index (XB) and partition entropy (PE) indicate a better clustering quality. The maximum for fuzzy SI and the minimums for XB and PE all suggested an optimal two-cluster solution. The grouped box-plot shows the between-cluster comparison results of the original values of the PANSS three subscales.3D visualization of the two resulting clusters was also presented.

B) The radar chart shows the six GMM clusters.

The red line depicts the median, the green diamond depicts the mean; For the grouped box-plot: The black dashed line depicts the median, the yellow diamond depicts the mean, and the whiskers represent the 5th and 95th percentiles.

Figure S21. Clustering the residuals of the three PANSS subscales after adjusting for age, gender, illness duration and total PANSS score



A) Internal validity indices used for determining the optimal cluster number. Higher values of fuzzy Silhouette index (SI), lower values of Xie and Beni index (XB) and partition entropy (PE) indicate a better clustering quality.

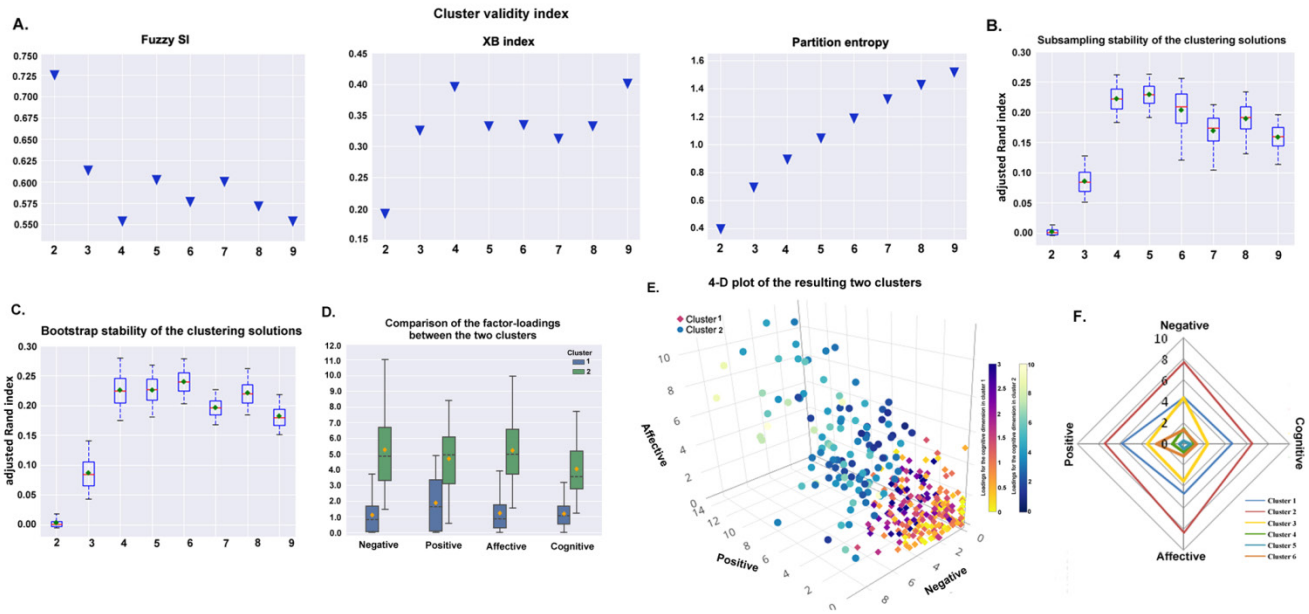
B-C) Clustering stability assessment based on subsampling and bootstrap resampling technique. The adjusted Rand index reflects the convergent assignment of the patient-pairs to the clusters between the sub-samples/bootstraps and the original sample.

D) The cluster number is set to 2 when creating the 3-D plot for a comparison with the optimal two clusters generated based on our four factors. The grouped box-plot shows the between-cluster comparison results of the residuals of the PANSS subscales.

E) Subsampled and bootstrapped values for cluster validity index: Higher values of fuzzy SI and lower values of XB and PE indicate a better clustering quality.

The red line depicts the median, the green diamond depicts the mean; For the grouped box-plot: the black dashed line depicts the median, the yellow diamond depicts the mean, and the whiskers represent the 5th and 95th percentiles.

Figure S22. Clustering the raw loadings on the four OPNMF factors without adjusting for age, gender, illness duration and total PANSS score



A) Internal validity indices used for determining the optimal cluster number. Higher values of fuzzy Silhouette index (SI), lower values of Xie and Beni index (XB) and partition entropy (PE) indicate a better clustering quality. The maximum for FSI and the minimums for XB and PE all suggested an optimal two-cluster solution.

B-C) Clustering stability assessment based on subsampling and bootstrap resampling technique. Adjusted Rand index reflects the convergent assignment of the patient-pairs to the clusters between the sub-samples/bootstraps and the original sample.

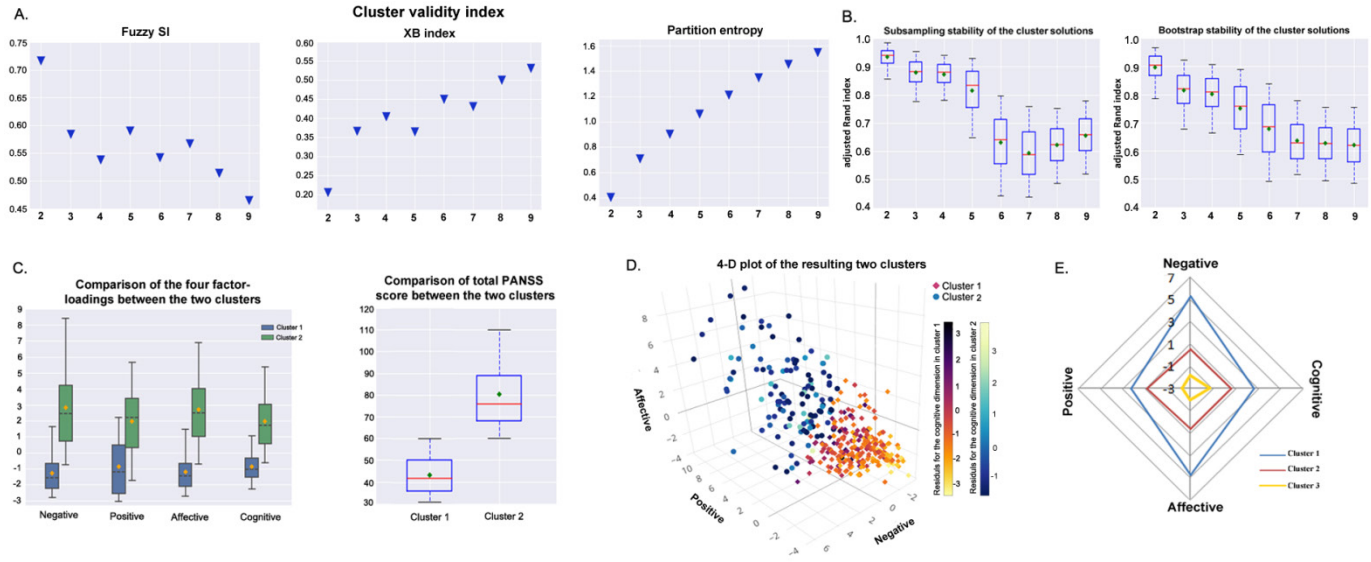
D) The grouped box-plot shows the between-cluster comparison results of the four factor-loadings.

E) 4D visualization of the two resulting clusters.

F) The radar chart shows the six GMM clusters.

The red line depicts the median, the green diamond depicts the mean; For the grouped box-plot: the black dashed line depicts the median, the yellow diamond depicts the mean, and whiskers represent the 5th and 95th percentiles.

Figure S23. Clustering the residuals of the loadings on the four OPNMF factors with adjusting for age, gender, illness duration, but without total PANSS score adjustment



A) Internal validity indices used for determining the optimal cluster number. Higher values of fuzzy Silhouette index (SI), lower values of Xie and Beni index (XB) and partition entropy (PE) indicate a better clustering quality. The maximum for fuzzy SI and the minimums for XB and PE all suggested an optimal two-cluster solution.

B) Clustering stability assessment based on subsampling and bootstrap resampling technique. Adjusted Rand index here reflects the convergent assignment of the patient-pairs to the clusters between the subsamples/ bootstraps and the original sample.

C) The grouped box-plot shows the between-cluster comparison results of the residuals of the four factor-loadings, and the box-plot shows the between-cluster comparison for total PANSS score.

D) 4D visualization of the two resulting clusters (total PANSS score was not adjusted).

E) The radar chart shows the three GMM clusters.

The red line depicts the median, the green diamond depicts the mean; For the grouped box-plot: the black dashed line depicts the median, the yellow diamond depicts the mean, and the whiskers represent the 5th and 95th percentiles.

Figure S24. Histograms show the head motion metric (DVARS) for the patients of core subtypes (left), and the all patients including those with ambiguous subtype memberships (right), in order to filter out the patients with excessive head motions. Ten DVARS units refer to 1% BOLD signal change

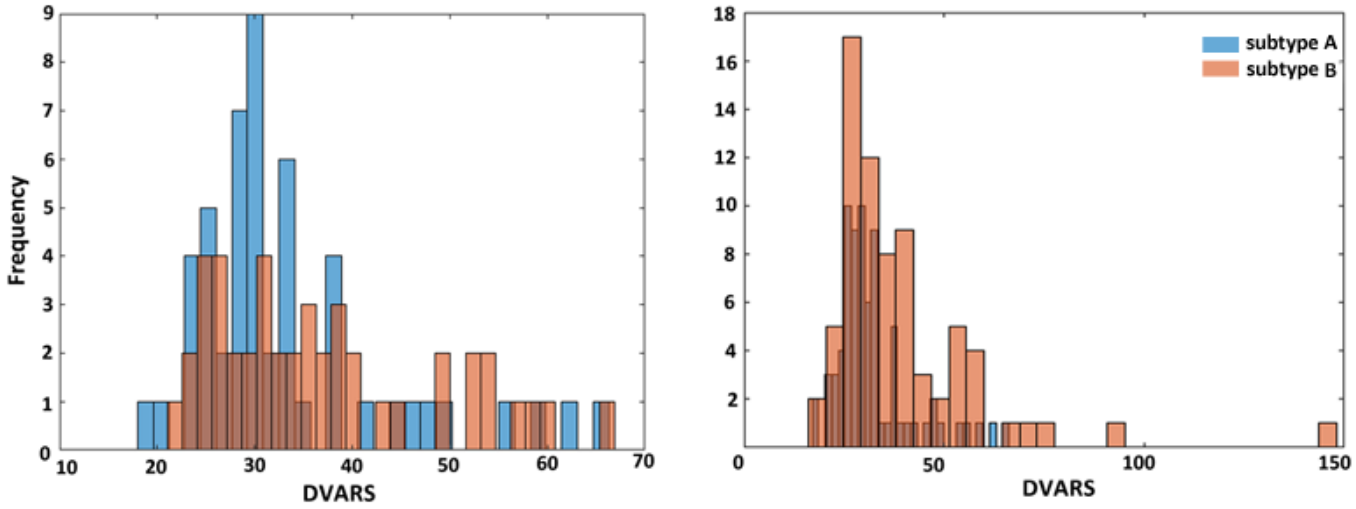


Figure S25. Schematic for 10-fold cross-validation in classification analysis

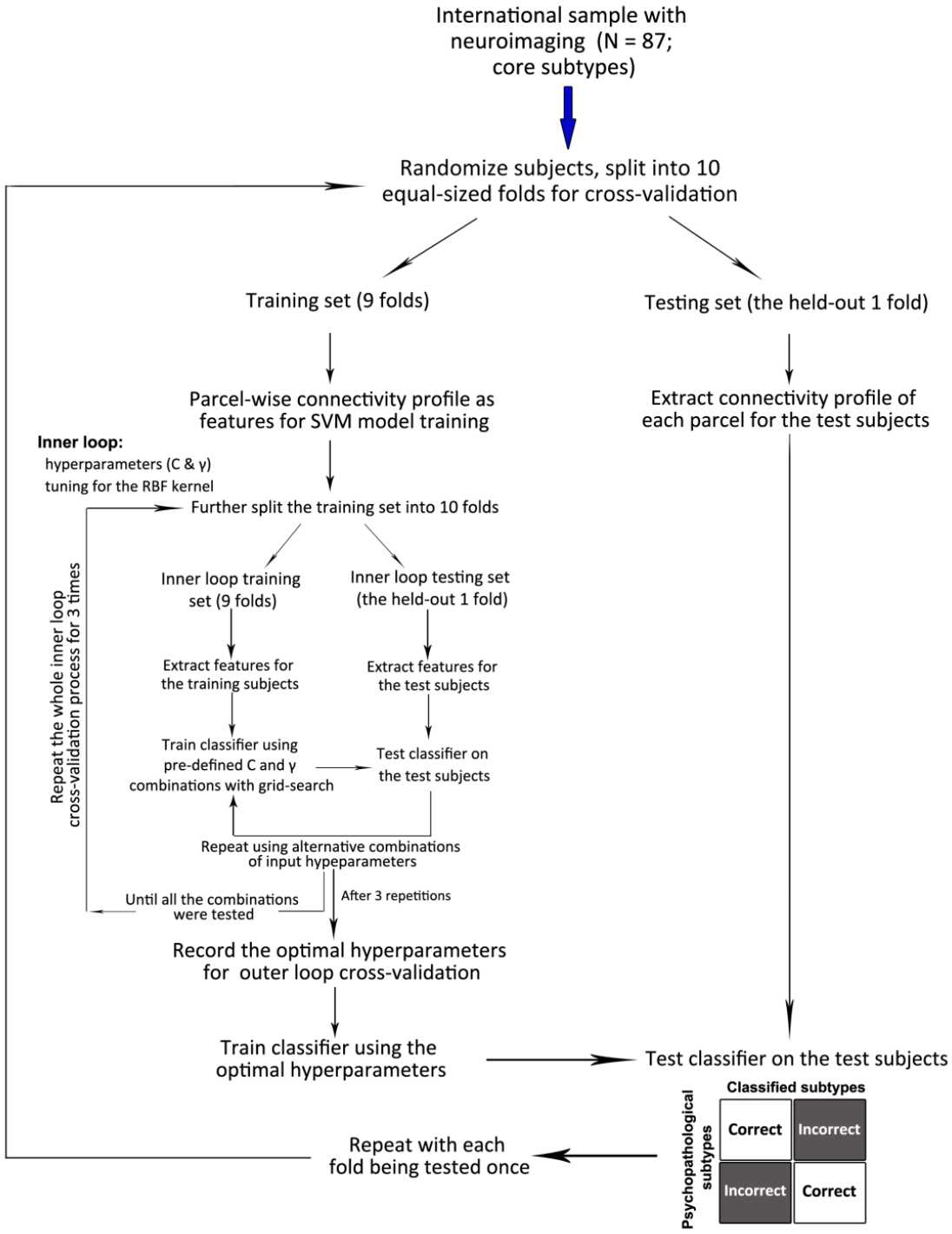
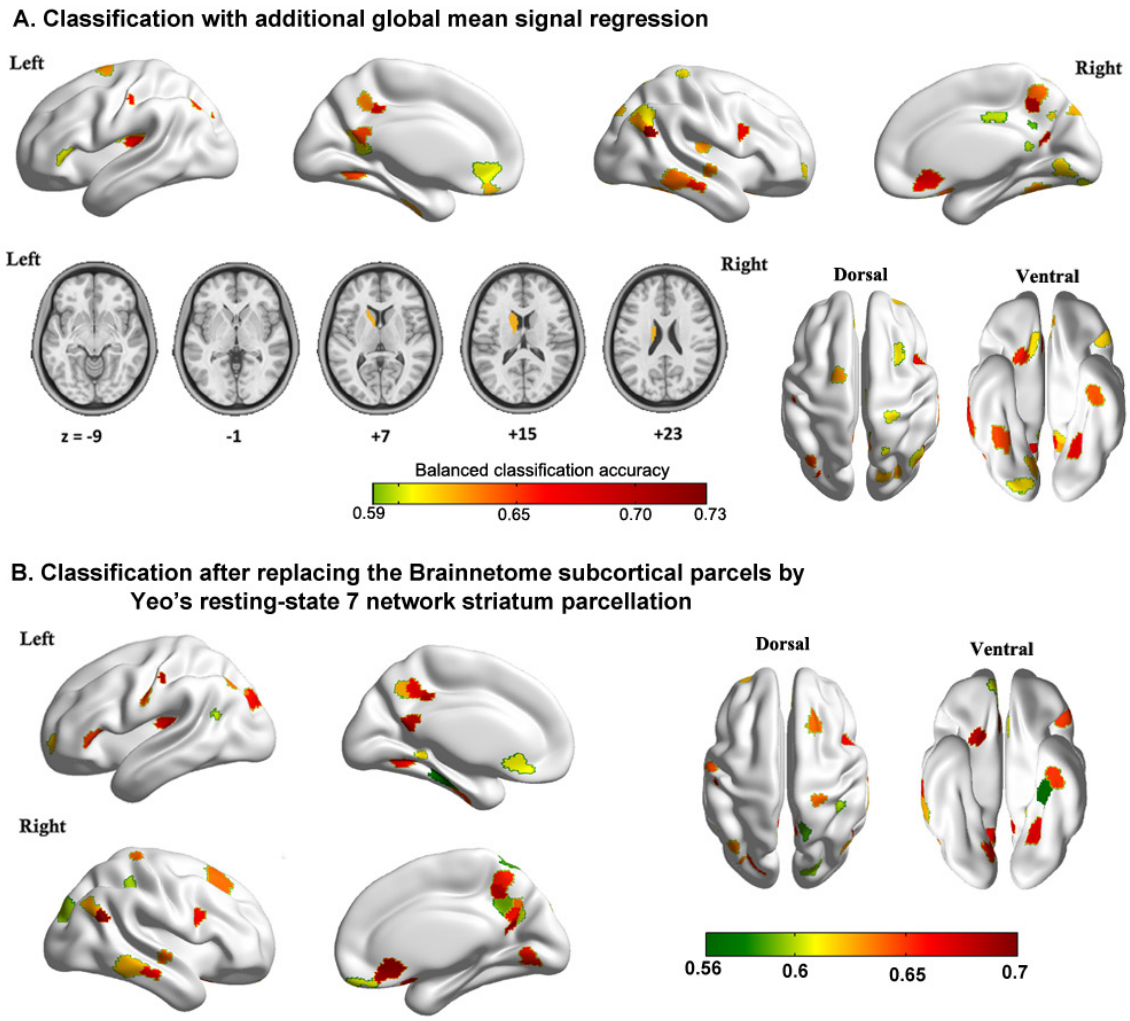


Figure S26. Classification of the two psychopathological subtypes from resting-state functional connectivity patterns with global mean signal removal (A), as well as after replacing the Brainnetome subcortical parcels by a resting-state functional connectivity derived striatum parcels (B)



Supplemental References:

1. Association AP (1994): Diagnostic and statistical manual of mental disorders (DSM-IV). American Psychiatry Association, Washington, DC.
2. Andreasen NC, Flaum M, Arndt S (1992): The Comprehensive Assessment of Symptoms and History (CASH): an instrument for assessing diagnosis and psychopathology. *Arch Gen Psychiatry* 49: 615-623.
3. Clos M, Diederen KM, Meijering AL, Sommer IE, Eickhoff SB (2014): Aberrant connectivity of areas for decoding degraded speech in patients with auditory verbal hallucinations. *Brain Struct Funct* 219: 581-594.
4. Chahine G, Richter A, Wolter S, Goya-Maldonado R, Gruber O (2017): Disruptions in the left frontoparietal network underlie resting state endophenotypic markers in schizophrenia. *Hum Brain Mapp* 38: 1741-1750.
5. Giel R, Nienhuis F (1996): SCAN-2.1: Schedules for clinical assessment in neuropsychiatry. Geneva/Groningen: WHO.
6. Vercammen A, Knegtering H, den Boer JA, Liemburg EJ, Aleman A (2010): Auditory hallucinations in schizophrenia are associated with reduced functional connectivity of the temporo-parietal area. *Biol Psychiatry* 67:912-918.
7. Spitzer RL, Gibbon ME, Skodol AE, Williams JB, First MB (2002): DSM-IV-TR casebook: A learning companion to the diagnostic and statistical manual of mental disorders, text rev, *American Psychiatric Publishing*, Inc.
8. Lefebvre S, Demeulemeester M, Leroy A, Delmaire C, Lopes R, Pins D, *et al.* (2016): Network dynamics during the different stages of hallucinations in schizophrenia. *Hum Brain Mapp* 37: 2571-2586.
9. Peters H, Riedl V, Manoliu A, Scherr M, Schwerthöffer D, Zimmer C, *et al.*(2017): Changes in extra-striatal functional connectivity in patients with schizophrenia in a psychotic episode. *Br J Psychiatry* 210: 75-82.
10. Sorg C, Manoliu A, Neufang S, Myers N, Peters H, Schwerthöffer D, *et al.* (2012): Increased intrinsic brain activity in the striatum reflects symptom dimensions in schizophrenia. *Schizophr Bull* 39: 387-395.
11. Mayer AR, Ruhl D, Merideth F, Ling J, Hanlon FM, Bustillo J, *et al.* (2013): Functional imaging of the hemodynamic sensory gating response in schizophrenia. *Hum Brain Mapp* 34: 2302-2312.
12. Association AP (2013): Diagnostic and statistical manual of mental disorders (DSM-5®), American Psychiatric Pub.
13. Regenbogen C, Kellermann T, Seubert J, Schneider DA, Gur RE, Derntl B, *et al.* (2015): Neural responses to dynamic multimodal stimuli and pathology-specific impairments of social cognition in schizophrenia and depression. *Br J Psychiatry* 206: 198-205.

14. Schilbach L, Derntl B, Aleman A, Caspers S, Clos M, Diederen KM, *et al.* (2016): Differential patterns of dysconnectivity in mirror neuron and mentalizing networks in schizophrenia. *Schizophr Bull* 42: 1135-1148.
15. First MB, Spitzer RL, Gibbon M, Williams JBW (1994): Structured Clinical Interview for DSM-IV Axis I Disorders-Patient Version (SCID-P). American Psychiatric Press.
16. Collinson SL, Gan SC, San Woon P, Kuswanto C, Sum MY, Yang GL, *et al.* (2014): Corpus callosum morphology in first-episode and chronic schizophrenia: combined magnetic resonance and diffusion tensor imaging study of Chinese Singaporean patients. *Br J Psychiatry* 204: 55-60.
17. Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, Kwiatkowski DJ, *et al.* (2016): Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genetics* 48: 600-606.
18. Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013): Network-based stratification of tumor mutations. *Nat Methods* 10: 1108-1115.
19. Sadanandam A, Lyssiotis CA, HomicskoK,Collisson EA, Gibb WJ, Wullschleger S, *et al.* (2013): A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med* 19: 619-625.
20. Sotiras A, Resnick SM, Davatzikos C (2015): Finding imaging patterns of structural covariance via non-negative matrix factorization. *NeuroImage* 108: 1-16.
21. Daubechies I, Roussos E, Takerkart S, Benharrosh M, Golden C, D'Ardenne K, *et al.*(2009): Independent component analysis for brain fMRI does not select for independence. *Proc Natl Acad Sci U S A* 106: 10415-10422.
22. Avants BB, Libon DJ, Rascovsky K, Boller A, McMillan CT, Massimo L, *et al.* (2014): Sparse canonical correlation analysis relates network-level atrophy to multivariate cognitive measures in a neurodegenerative population. *Neuroimage* 84: 698-711.
23. Lee DD, Seung HS (2001): Algorithms for non-negative matrix factorization. Paper presented at: Advances in neural information processing systems.
24. Yang Z, Oja E (2010): Linear and nonlinear projective nonnegative matrix factorization. *IEEE T Neural Networ* 21: 734-749.
25. Boutsidis C, Gallopoulos E (2008): SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognit* 41: 1350-1362.
26. Brunet J-P, Tamayo P, Golub TR, Mesirov JP (2004): Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* 101: 4164-4169.

27. Jaccard P (1901): Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat* 37: 547-579.
28. Hubert L, Arabie P (1985): Comparing partitions. *J Classif* 2: 193-218.
29. Meilă M (2007): Comparing clusterings-an information based distance. *J Multivar Anal* 98: 873-895.
30. Kahnt T, Chang LJ, Park SQ, Heinzle J, Haynes JD (2012): Connectivity-based parcellation of the human orbitofrontal cortex. *J Neurosci* 32: 6240-6250.
31. Raguideau S, Plancade S, Pons N, Leclerc M, Laroche B (2016): Inferring Aggregated Functional Traits from Metagenomic Data Using Constrained Non-negative Matrix Factorization: Application to Fiber Degradation in the Human Gut Microbiota. *PLoS Comput Biol* 12: e1005252.
32. Leys C, Ley C, Klein O, Bernard P, Licata L (2013): Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol* 49: 764-766.
33. Whittaker JR, Driver ID, Bright MG, Murphy K (2016): The absolute CBF response to activation is preserved during elevated perfusion: Implications for neurovascular coupling measures. *NeuroImage* 125: 198-207.
34. Wimber M, Alink A, Charest I, Kriegeskorte N, Anderson MC (2015): Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nat Neurosci* 18: 582-589.
35. Miller J (1991): Reaction time analysis with outlier exclusion: Bias varies with sample size. *Q J Exp Psychol A* 43: 907-912.
36. Rousseeuw PJ, Croux C (1993): Alternatives to the median absolute deviation. *J Am Stat Assoc* 88: 1273-1283.
37. Kay SR, Sevy S (1990): Pyramidical model of schizophrenia. *Schizophr Bull* 16: 537-545.
38. Emsley R, Rabinowitz J, Torreman M, Group R-I-EPGW (2003): The factor structure for the Positive and Negative Syndrome Scale (PANSS) in recent-onset psychosis. *Schizophr Res* 61: 47-57.
39. Van den Oord EJ, Rujescu D, Robles JR, Giegling I, Birrell C, JózsefBukszár, *et al.* (2006): Factor structure and external validity of the PANSS revisited. *Schizophr Res* 82: 213-223.
40. van der Gaag M, Cuijpers A, Hoffman T, Remijsen M, Hijman R, de Haan L, *et al.*(2006): The five-factor model of the Positive and Negative Syndrome Scale I: confirmatory factor analysis fails to confirm 25 published five-factor solutions. *Schizophr Res* 85: 273-279.
41. White L, Harvey PD, Opler L, Lindenmayer J (1997): Empirical assessment of the factorial structure of clinical symptoms in schizophrenia. *Psychopathology* 30: 263-274.42.
42. Kim JH, Kim S-Y, Lee J, Oh KJ, Kim YB, Cho ZH (2012): Evaluation of the factor structure of symptoms in

patients with schizophrenia. *Psychiatry Res* 197: 285-289.

43. Cronbach LJ (1951): Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297-334.
44. Bentler PM (1985): Efficient estimation via linearization in structural models. *Multivar Anal*, Amsterdam, Elsevier, pp 9-42.
45. Chou CP, Bentler PM (1995): Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues, and applications. Thousand Oaks, CA, Sage, pp 37-54.
46. Kline RB (2015): Principles and practice of structural equation modeling. Guilford publications, New York.
47. Marsh HW, Hau K-T, Wen Z (2004): In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in over generalizing Hu and Bentler's (1999) findings. *Struct Equ Model* 11: 320-341.
48. Bezdek JC (1981): Objective function clustering. In: Pattern recognition with fuzzy objective function algorithms. Springer; pp 43-93.
49. Fadili M-J, Ruan S, Bloyet D, Mazoyer B (2001): On the number of clusters and the fuzziness index for unsupervised FCA application to BOLD fMRI time series. *Med Image Anal* 5: 55-67.
50. Tan T, Choi JY, Hwang H (2015): Fuzzy clusterwise functional extended redundancy analysis. *Behaviormetrika* 42: 37-62.
51. Ozkan I, Turksen I (2007): Upper and lower values for the level of fuzziness in FCM. in Fuzzy Logic, Springer; pp 99-112.
52. Campello RJ, Hruschka ER (2006): A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets Syst* 157: 2858-2875.
53. Xie XL, Beni G (1991): A validity measure for fuzzy clustering. *IEEE T Pattern Anal* 13: 841-847.
54. Rousseeuw PJ (1987): Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20: 53-65.
55. Ben-Hur A, Guyon I (2003): Detecting stable clusters using principal component analysis. *Funct Genomics Methods Protoc* 224: 159-182.
56. Fraley C, Raftery AE (2002): Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 97: 611-631.
57. Scrucca L, Fop M, Murphy TB, Raftery AE (2016): mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R journal* 8: 289.
58. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE (2012): Spurious but systematic correlations in

functional connectivity MRI networks arise from subject motion. *Neuroimage* 59: 2142-2154.

59. Power JD, Schlaggar BL, Petersen SE (2015): Recent progress and outstanding issues in motion correction in resting state fMRI. *Neuroimage* 105: 536-551.
60. Parker D, Liu X, RazlighiQR (2017): Optimal slice timing correction and its interaction with fMRI parameters and artifacts. *Med Image Anal* 35: 434-445.
61. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012): Fsl. *Neuroimage* 62: 782-790.
62. Calhoun VD, Wager TD, Krishnan A, Rosch KS, Seymour KE, Nebel MB, *et al.* (2017): The impact of T1 versus EPI spatial normalization templates for fMRI data analyses. *Hum Brain Mapp* 38: 5331-5342.
63. Ashburner J, Friston KJ (2011): Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. *NeuroImage* 55: 954-967.
64. Varikuti DP, Hoffstaedter F, Genon S, Schwender H, Reid AT, Eickhoff SB (2017): Resting-state test–retest reliability of a priori defined canonical networks over different preprocessing steps. *Brain Struct Funct* 222: 1447-1468.
65. Li J, Kong R, Liegeois R, *et al.* (2019): Global signal regression strengthens association between resting-state functional connectivity and behavior. *NeuroImage* 196: 126-141.
66. Murphy K, Fox MD. (2017): Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage* 154: 169-173.
67. MurphyK, Birn RM, Handwerker DA, Jones TB, Bandettini PA. (2009): The impact of global signal regression on resting state correlations: areanti-correlated networks introduced? *NeuroImage* 44: 893-905.
68. Schaefer A, Kong R, Gordon EM, Laumann T, ZuoXN, HolmesA, *et al.* (2017): Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb Cortex* 28: 3095-3114.
69. Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, *et al.*(2016): The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cereb Cortex* 26: 3508-3526.
70. Shen X, Tokoglu F, Papademetris X, Constable RT. (2013): Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage* 82:403–415.
71. Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS. (2012): A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum Brain Mapp* 33:1914–1928
72. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, *et al.*(2016): A multi-modal parcellation of human cerebral cortex. *Nature* 536: 171-178
73. Gordon EM, Laumann TO, Adeyemo B, HuckinsJF, Kelley WM, Petersen SE. (2016): Generation and

evaluation of a cortical area parcellation from resting-state correlations. *Cereb Cortex* 26:288–303.

74. Choi EY, Yeo BTT, Buckner RL. (2012): The organization of the human striatum estimated by intrinsic functional connectivity. *J Neurophysiol* 108(8): 2242-2263.
75. Snoek L, Miletic S, Scholte HS (2009): How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage* 184: 741-760.
76. BrodersenKH, Ong CS, Stephan KE, Buhmann JM (2010): The balanced accuracy and its posterior distribution. Paper presented at: Pattern Recognit (ICPR), 20th international conference on 2010.
77. Golland P, Fischl B (2003): Permutation tests for classification: towards statistical significance in image-based studies. Paper presented at: Biennial International Conference on Inf Process Med Imaging.
78. Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M (2005): Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *NeuroImage* 28: 980-995.
79. Laird AR, Eickhoff SB, Kurth F, Fox PM, Uecker AM, Turner JA, *et al.*(2009): ALE meta-analysis workflows via the brainmap database: progress towards a probabilistic functional brain atlas. *Front Neuroinformatics* 3:23.
80. Cieslik EC, Zilles K, Caspers S, Roski C, Kellermann TS, Jakobs O, *et al.* (2012): Is there “one” DLPFC in cognitive action control? Evidence for heterogeneity from co-activation-based parcellation. *Cereb Cortex* 23:2677-2689.
81. Clos M, Amunts K, Laird AR, Fox PT, Eickhoff SB (2013): Tackling the multifunctional nature of Broca's region meta-analytically: co-activation-based parcellation of area 44. *Neuroimage* 83:174-188.
82. Genon S, Li H, Fan L, Müller VI, Cieslik EC, Hoffstaedter F, *et al.* (2017): The right dorsal premotor mosaic: organization, functions, and connectivity. *Cereb Cortex* 27:2095-2110.
83. Rottschy C, Caspers S, Roski C, ReetzI K, Dogan I, Schulz JB, *et al.* (2013): Differentiated parietal connectivity of frontal regions for “what” and “where” memory. *Brain Struct Funct* 218: 1551-1567.