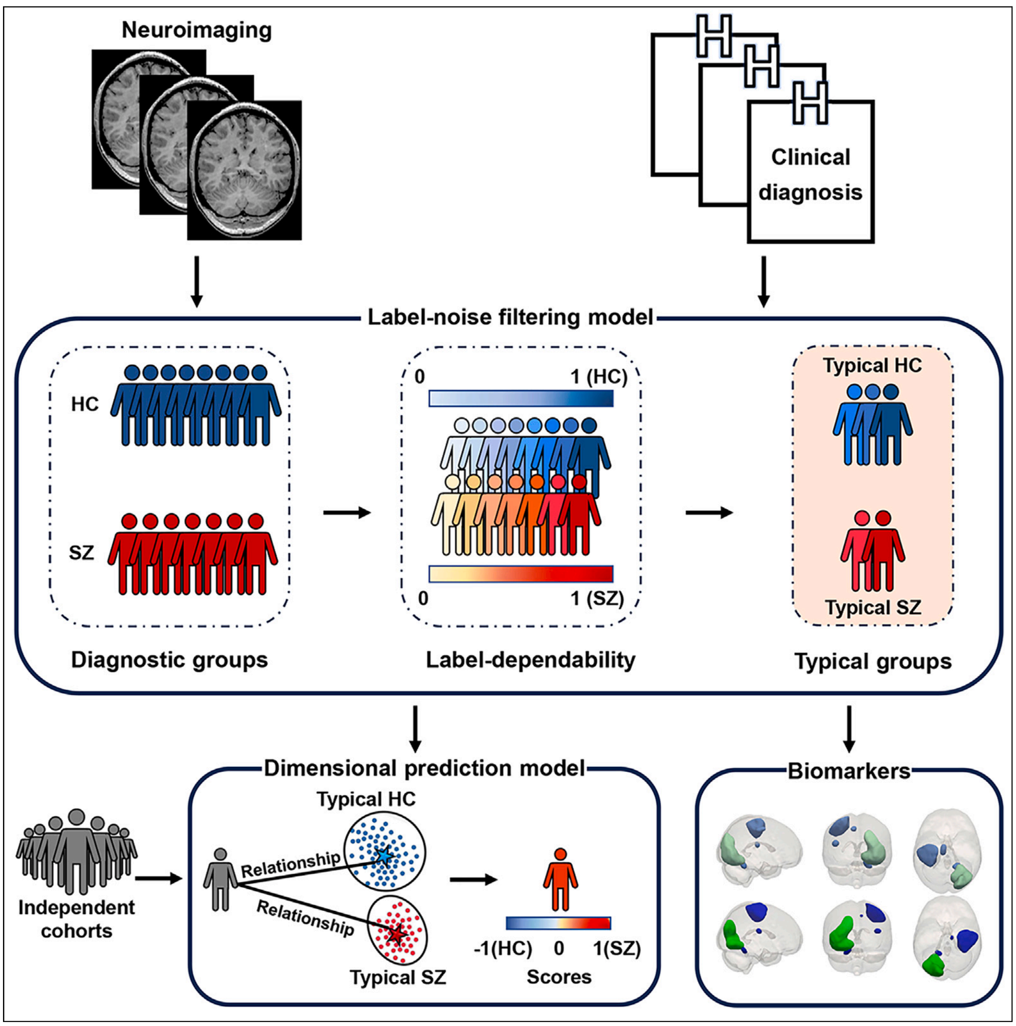
iScience

CellPress
OPEN ACCESS

Article

More reliable biomarkers and more accurate prediction for mental disorders using a label-noise filtering-based dimensional prediction method



Ying Xing, Theo G.M. van Erp, Godfrey D. Pearlson, Peter Kochunov, Vince D. Calhoun, Yuhui Du

duyuhui@sxu.edu.cn

Highlights

We propose a neuroimaging-based method for dimensional predictions of mental disorders

It utilizes a label-noise filtering strategy to accurately identify typical subjects

It can output a continuous score indicating the pathology for mental disorders

It facilitates the identification of reliable biomarkers for mental disorders

Xing et al., iScience *27*, 109319
March 15, 2024 © 2024 The Author(s).
https://doi.org/10.1016/j.isci.2024.109319

iScience

CellPress OPEN ACCESS

Article

More reliable biomarkers and more accurate prediction for mental disorders using a label-noise filtering-based dimensional prediction method

Ying Xing,[1] Theo G.M. van Erp,[2,3] Godfrey D. Pearlson,[4,5] Peter Kochunov,[6] Vince D. Calhoun,[7] and Yuhui Du[1,8,*]

SUMMARY

The integration of neuroimaging with artificial intelligence is crucial for advancing the diagnosis of mental disorders. However, challenges arise from incomplete matching between diagnostic labels and neuroimaging. Here, we propose a label-noise filtering-based dimensional prediction (LAMP) method to identify reliable biomarkers and achieve accurate prediction for mental disorders. Our method proposes to utilize a label-noise filtering model to automatically filter out unclear cases from a neuroimaging perspective, and then the typical subjects whose diagnostic labels align with neuroimaging measures are used to construct a dimensional prediction model to score independent subjects. Using fMRI data of schizophrenia patients and healthy controls (n = 1,245), our method yields consistent scores to independent subjects, leading to more distinguishable relabeled groups with an enhanced classification accuracy of 31.89%. Additionally, it enables the exploration of stable abnormalities in schizophrenia. In summary, our LAMP method facilitates the identification of reliable biomarkers and accurate diagnosis of mental disorders using neuroimages.

INTRODUCTION

Mental disorders are associated with brain functional and structural impairments and affect about 970 million people worldwide.[1] The traditional case (i.e., patients with mental disorders) versus control (i.e., healthy subjects) diagnosis provides binary classifications, where a subject is assigned to either case or control category according to assessments of clinical phenomenology.[2] Despite the abundance of diagnostic rating scales used to assess subjects' symptoms from multiple aspects for diagnosing mental disorders, this subjective and categorical manner may result in subjects with label noise whose diagnostic labels are inconsistent with the underlying brain abnormalities.[3,4] The sources of label noise may include: (1) insufficient information: diagnosis solely based on clinical manifestations may fail to capture underlying brain functional and structural abnormalities, (2) subjectivity of diagnosis: different experts may produce different labeling results due to variations in expertise, and (3) problems in understanding and communication: psychiatric diagnosis relying on self-reporting by the subjects may be unreliable.[5,6] The source of label noise is closely related to the type of mental disorders. For example, the second and third sources of label noise are more common in anxiety disorders than in schizophrenia (SZ). In addition, the heterogeneity of mental disorders also makes it easy to generate label noise that does not match the neuroimaging. Prediction models based on heterogeneous groups of patients may reveal heterogeneous patterns of brain structural or functional changes.[7] In short, diagnoses for mental disorders based on clinical data do not completely match neuroimaging, and the possible inconsistency between diagnosis labels and neuroimaging measures may affect the validity of exploring reliable biomarkers and constructing accurate prediction models for diagnosing mental disorders.

With the development of massive neuroimaging data and advanced machine learning, numerous studies aim to explore biomarkers and construct prediction models for recognizing patients with various mental disorders.[2,8] Supervised learning focuses on detecting biomarkers and constructing neuroimaging-based classifiers for mental disorders with the guidance of diagnostic labels.[9,10] Although promising biomarkers and prediction models have been developed, the reliability and generalizability of the outcomes remain controversial.[2,11–14] One possible explanation is the potential inconsistency between the diagnostic labels and neuroimaging measures, which may mislead the supervised classification and hinder the generation of stable and reliable outcomes.[3,15] Considering the existence of the potential inconsistency, data-driven clustering methods that ignore all labels and cluster subjects into homogeneous groups have recently received great

[1]School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
[2]Department of Psychiatry and Human Behavior, School of Medicine, University of California, Irvine, Irvine, CA 92617, USA
[3]Center for the Neurobiology of Learning and Memory, University of California, Irvine, Irvine, CA 92617, USA
[4]Departments of Psychiatry and of Neurobiology, Yale University, New Haven, CT 06519, USA
[5]Olin Neuropsychiatry Research Center, Institute of Living, Hartford, CT 06106, USA
[6]Maryland Psychiatric Research Center and Department of Psychiatry, University of Maryland, School of Medicine, Baltimore, MD 21201, USA
[7]Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA 30030, USA
[8]Lead contact
*Correspondence: duyuhui@sxu.edu.cn
https://doi.org/10.1016/j.isci.2024.109319

attention.[16–19] Although unsupervised clustering methods provide a promising strategy to redefine nosology, the stability, interpretability, and reproducibility of the outcomes have been questioned.[20,21] One possible reason for the skepticism is that a lack of domain knowledge provided by expert guidance (i.e., diagnostic labels) may result in inconsistent and inexplicable results.[10,22] Fortunately, in the field of machine learning, there are numerous well-established approaches focused on eliminating inaccurately labeled samples and utilizing valuable information to build more reliable models.[23–28] Efforts are also underway to apply noise-cleansing techniques in the mental health field, aiming to reduce the inconsistency and enhance the reliability of findings in the field. There has been a novel attempt using structural magnetic resonance imaging (MRI) measures to mitigate the negative impact caused by the inconsistency rather than discarding all labels indiscriminately.[29] Subjects were used to build multiple support vector machine (SVM) classifiers to relabel subjects that were unanimously mislabeled by all classifiers. The process was repeated using the subjects with refreshed labels until the number of mislabeled subjects fell below a given threshold. However, the reliability of classifiers built on the subjects with refreshed labels from the previous iteration may greatly affect the reliability of subsequent label refreshing. In short, it remains a challenge to decrease the inconsistency between diagnostic labels and neuroimaging measures in order to explore reliable biomarkers and facilitate accurate predictions for mental disorders.

Furthermore, the current categorical diagnostic methods are also controversial, and a dimensional perspective seems more appropriate for the diagnoses of mental disorders.[30,31] Many mental disorders occur along a continuum from mild to severe, and healthy controls (HCs) are also at varying risk of developing mental disorders.[32,33] Research found that mental disorders exhibit a continuous pattern in brain function and structure using neuroimaging, which is associated with the risk of the disorders.[34,35] The initiative to integrate clinical diagnostic information with neuroimaging of mental disorders has been proposed to reflect the severity of subjects in behavior, cognition, and other aspects, aiming to better assist in dimensional diagnosis, treatment, and prognosis.[36] However, it remains challenging to utilize different levels of data that are related but not directly corresponding in studying mental disorders. In brief, it is important to identify subjects with matching diagnostic labels and neuroimaging measures to develop dimensional approaches that capture continuous changes in the brain associated with mental disorders.

To help explore more reliable biomarkers and benefit more accurate prediction for mental disorders, we propose a label-noise filtering-based dimensional prediction (LAMP) method using neuroimaging data. In our method, based on random forest, the inherent data structure and diagnostic labels are utilized automatically to select typical subjects that have more consistency between diagnostic labels and neuroimaging measures. Typical subjects with enhanced inter-group separability and intra-group compactness are then served as the benchmark to build a dimensional prediction model, which enables a quantitative analysis for the unseen (independent) subject to indicate the degree of pathology. Importantly, these clearly separable subjects can help identify stable biomarkers across different datasets from different sites. In this work, the reliability and generalizability of our method are validated using large-scale functional MRI (fMRI) data of SZ patients and HCs from four datasets. Thanks to the integration of clinical information and neuroimaging measures, the proposed method provides stable dimensional scores and identifies significant and consistent inter-site group differences, elucidating the underlying mechanisms for mental disorders. It is worth noting that this framework can be applied to other data modalities and multimodal data.

RESULTS

Typical subjects with enhanced intra-group compactness and inter-group separability from a neuroimaging perspective are identified

The overall workflow of the present work is shown in Figure 1. To evaluate our LAMP method, we employed the resting-state fMRI data of large-sample HCs and SZ patients (708 HCs and 537 SZ patients) from four datasets, including BSNIP (Bipolar and Schizophrenia Network on Intermediate Phenotypes), FBIRN (Function Biomedical Informatics Research Network), MPRC (Maryland Psychiatric Research Center), and COBRE (Centers for Biomedical Research Excellence) (see Table S1 for demographic information in detail). As shown in Figure 1A, we first extracted functional network connectivity (FNC) features for each subject using fMRI data via the NeuroMark, a fully automated independent component analysis pipeline.[37] It is worth noting that the nuisance effects including age, gender, head motion, and site effects were removed from the extracted FNC features so that more reliable outcomes can be obtained. In order to verify the result reproducibility, we employed a strict leave-one-dataset-out cross-validation procedure (see Figure 1B), whereby each of the four datasets successively served as the independent dataset for the evaluation and the remaining three as the source datasets for constructing the dimensional prediction model. As shown in Figure 1C, typical HCs and SZ subjects were then identified by building a complete random forest (CRF)-based label-noise filtering model[38] from each source dataset. Subjects that are close in feature space should have similar labels.[39] Thus, the CRF-based model determines the label-dependability of each subject by evaluating the heterogeneity levels among their surrounding subjects. This enables the identification of typical subjects who are predominantly surrounded by others of the same category. It is worth mentioning that we validated the ability of the CRF-based model in accurately identifying typical samples on three public datasets (see Table S3). Given a label-dependability threshold γ, we can obtain typical subjects with various consistency levels between diagnostic labels and neuroimaging measures. Increasing the threshold results in greater consistency between the labels and fMRI measures of the selected typical subjects. The number of typical subjects retained is linked to the quality of a dataset, which is influenced by factors such as the participants, the data collection equipment, and the expertise level of the annotators. Compared with the original groups containing all subjects with diagnostic labels in each source dataset, the identified typical HC and SZ groups would show greater intra-group compactness and inter-group separability using FNC features, indicating a strong consistency between their diagnostic labels and fMRI measures.

As shown in Figure 1E, intra-group compactness and inter-group separability of typical groups derived from each source dataset were evaluated from various aspects, including investigating inter-group differences, classification performance, separation between groups,

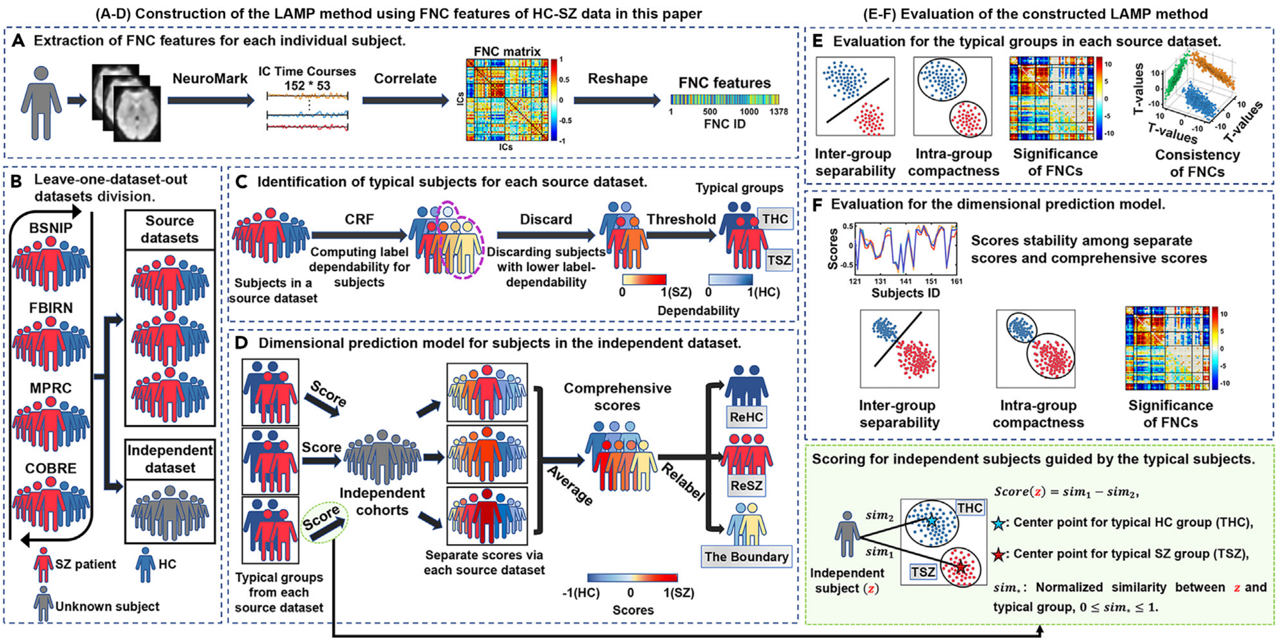


Figure 1. The overall workflow of the proposed label-noise filtering-based dimensional prediction (LAMP) method using fMRI data of HC-SZ

The construction of the LAMP method using FNC features of HC-SZ data is displayed in (A–D), and the evaluation of the method is displayed in (E and F). (A) We extracted the FNC features for each subject from four different fMRI datasets based on previous work, i.e., NeuroMark, which is a fully automated independent component analysis pipeline. (B) We sequentially took one of the four datasets as the independent dataset and the remaining three as the source datasets, namely, the leave-one-dataset-out division strategy. (C) We identified typical subjects in each source dataset. Specifically, we evaluated each subject's label-dependability that indicates the consistency level between the diagnostic label and fMRI measures via the CRF-based label-noise filtering model, and discarded subjects whose label-dependability was lower than a predefined threshold γ. After that, we thresholded the label-dependability of the remaining subjects, resulting in the typical subjects (i.e., typical HCs and typical SZ patients). Notably, the CRF-based model regards the subjects that are surrounded by the homogeneous subjects as typical subjects. (D) We constructed a dimensional prediction model guided by the typical subjects from various source datasets to provide a dimensional score for each independent subject. In detail, we first predicted a separate score revealing the degree of brain dysfunction for each independent subject according to its relationship to different typical groups in each source dataset. Next, we averaged the three separate scores derived from the source datasets to get a comprehensive score for the independent subject. Then, we relabeled the independent subjects according to the comprehensive score with an adaptive parameter τ and obtained the relabeled HC group, relabeled SZ group, and the Boundary group in which these subjects were mild and could not be categorized into HC or SZ group with enough confidence. (E) We verified the performance of the typical groups under different label-dependability thresholds γ in light of the inter-group separability and intra-group compactness and analyzed the significance and consistency of functional abnormality within the typical SZ group across multiple source datasets. (F) We evaluated the stability of the scores, intra-group compactness, and inter-group separability of the relabeled HC and SZ groups and analyzed the significant differences between the two relabeled groups within the independent dataset. IC denotes the independent component. THC, TSZ, ReHC, and ReSZ represent typical HC, typical SZ, relabeled HC, and relabeled SZ groups in a dataset, respectively. CRF is short for the complete random forest-based label-noise filtering model.

and cohesion within groups via widely used evaluation metrics, as well as visualization technology. Results provide evidence of the validity of identified typical subjects. Due to limited space, we thoroughly display the results obtained from taking the BSNIP dataset as the independent dataset and the remaining three (i.e., the FBIRN, MPRC, and COBRE datasets) as the source datasets for illustration. The results of taking FBIRN, MPRC, or COBRE dataset as the independent dataset are shown in supplemental information. Figures 2A−2C display the distribution of p values and the number of significant FNC features ($p < 0.01$ with Bonferroni-corrected two-sample t-test) in detecting group differences using statistical analysis for both the original and typical groups (under different label-dependability thresholds) in each source dataset. The results reveal more significant inter-group differences in FNC features between typical HC and SZ groups than between the original HC and SZ groups. It should be pointed out that the improved inter-group differences within the typical groups were not related to nuisance variables, such as age, gender, and head motion, measured by statistical analysis (Figures S2–S5). Besides, the classification ability in distinguishing HCs and SZ patients based on five classifiers was evaluated using a 5-fold cross-validation procedure for both original groups and typical groups (under different label-dependability thresholds) in each source dataset. Figures 2D–2F show that the classification accuracy of the typical groups increased by 30.5% on average across various classifiers compared with that of the original groups. Taking the FBIRN dataset (Figure 2D) as an example, the classifiers separated the typical groups with accuracies of over 80%, while separating the original groups was more difficult. In particular, four in the five classifiers achieved 100% accuracy on typical groups when the label-dependability threshold was 1, indicating improved inter-group separability of the typical groups. In addition, we visualized the two-dimensional (2D) projection of the intra- and inter-group structures intuitively for original groups and typical groups in the source datasets via t-SNE (stochastic neighbor

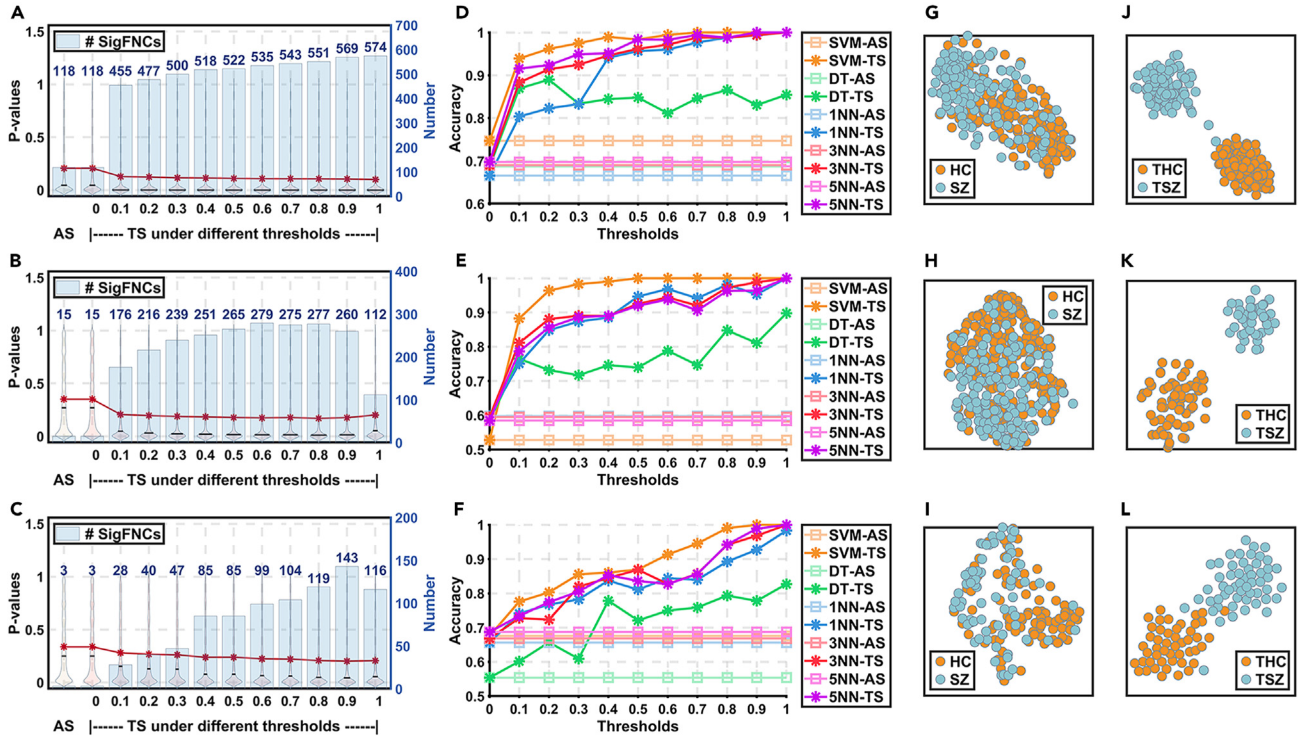


Figure 2. Evaluation of inter-group separability and intra-group compactness within original groups and within typical groups in the three source datasets, including FBIRN, MPRC, and COBRE datasets

(A–C) Distribution of p values and the number of significant FNC features ($p < 0.01$ with Bonferroni correction) for original groups and for the typical groups (under different label-dependability thresholds) in FBIRN, MPRC, and COBRE datasets, respectively. The horizontal axis represents the original groups containing all subjects with diagnostic labels (AS) and typical subjects (TS) at different label-dependability thresholds. The black vertical axis (left) represents the p values of FNC features between groups. The blue vertical axis (right) represents the number of significant FNC features ($p < 0.01$ with Bonferroni correction) between groups. It should be pointed out that when the threshold is 0, typical subjects are equivalent to the original subjects without label-noise filtering. It is shown that the number of significant FNC features between typical groups is several times higher than that between original groups.

(D–F) Average classification accuracy on original groups and on the typical groups (under different label-dependability thresholds) in FBIRN, MPRC, and COBRE datasets, respectively. The used five classifiers include support vector machine (SVM), decision tree (DT), 1-nearest neighborhood classifier (1NN), 3-nearest neighborhood classifier (3NN), and 5-nearest neighborhood classifier (5NN). Similarly, the horizontal axis represents the typical subjects at different label-dependability thresholds, and the typical subjects are equivalent to the original subjects when the threshold is 0. For the same classifier, the classification accuracy of typical groups is superior to that of original groups.

(G–I) 2D projection for original groups in FBIRN, MPRC, and COBRE datasets, respectively.

(J–L) 2D projection for typical groups identified under the 0.8 label-dependability threshold in FBIRN, MPRC, and COBRE datasets, respectively. The typical groups (THC vs. TSZ) are more separable than the original groups (HC vs. SZ). AS and TS represent original groups containing all subjects with diagnostic labels and typical subjects in the source dataset, respectively. THC and TSZ represent typical HC and SZ groups, respectively. # SigFNCs represents the number of significant FNC features ($p < 0.01$ with Bonferroni correction).

embedding) technology[40] in Figures 2G–2L. These figures show that typical subjects were grouped more clearly than all subjects in each source dataset, manifesting improved intra-group compactness within the typical groups. Furthermore, we found that the identified typical groups had better intra-group compactness and inter-group separability than the original groups by comparing five widely used evaluation metrics (Table S4). While using other datasets (FBIRN, MPRC, or COBRE) as the independent dataset, the results (Figure S6; Table S4) also support that the identified typical subjects from source datasets presented greater intra-group compactness and inter-group separability than the original groups. The increased separability between typical groups in various datasets further demonstrated the capability of the CRF-based model in accurately identifying subjects whose labels align with fMRI measures. In brief, substantial experiments highlight the enhanced intra-group compactness and inter-group separability of typical subjects in the source datasets. This demonstrates a marked consistency between their labels and fMRI measures, affirming their reliability for model construction.

Typical groups exhibit more significant and consistent inter-group differences across multiple source datasets

To reveal impaired brain function in SZ patients using typical subjects, we investigated significantly different FNC features between the typical HC group and the typical SZ group, and validated their consistency across multiple source datasets. Here, due to limited space, we only show the significant FNC group differences identified in FBIRN, MPRC, and COBRE datasets for the condition of taking BSNIP data as the

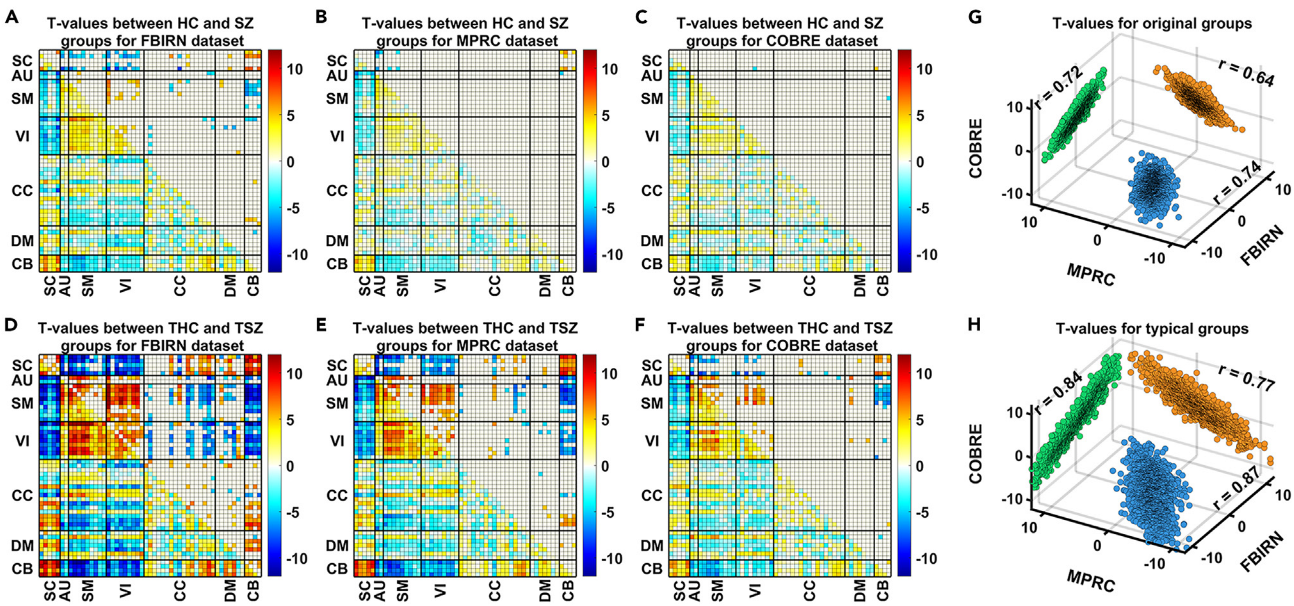


Figure 3. Inter-group differences for original groups and typical groups in the source datasets, including FBIRN, MPRC, and COBRE datasets

(A–C) Two-sample t-test T-values of FNCs (upper triangle: $p < 0.01$ with Bonferroni correction, lower triangle: no correction) using original groups (HC vs. SZ) in the source datasets.

(D–F) Two-sample t-test T-values of FNCs (upper triangle: $p < 0.01$ with Bonferroni correction, lower triangle: no correction) using typical groups (THC vs. TSZ) identified from the source datasets. In (A–F), 53 intrinsic connectivity networks are divided into seven brain functional domains, including sub-cortical (SC), auditory (AU), sensorimotor (SM), visual (VI), cognitive-control (CC), default-mode (DM), and cerebellar (CB) domains. The inter-group differences in the typical groups are more noticeable than those in the original groups. THC and TSZ represent typical HC and SZ groups, respectively.

(G) Consistency of T-values between any two source datasets using the original groups.

(H) Consistency of T-values between any two source datasets using the typical groups. The x axis, y axis, and z axis in (G–H) represent the T-values of FNC features in different datasets. Symbol r represents the correlation coefficient between T-values of FNC features in paired source datasets. In (G and H), the typical groups show more consistent inter-group differences than the original groups.

independent dataset. T-values of FNCs were calculated via two-sample t-tests across typical groups in each source dataset to evaluate the significance of the inter-group differences. Pearson correlation coefficients of the T-values in any two source datasets were calculated to evaluate the consistency of the inter-group differences. For comparison, the significance and consistency of the inter-group differences in FNC across multiple sources datasets were also computed for the original groups. Compared with the original groups (Figures 3A–3C), the T-values of significant FNC features ($p < 0.01$ with or without Bonferroni correction) between typical groups (Figures 3D–3F) support a more prominent inter-group difference. As shown in Figures 3G–3H, across source datasets, typical groups present more consistent inter-group differences than original groups (mean correlation coefficient $r = 0.83$ versus $r = 0.7$). That means we found more prominent and consistent inter-group differences in FNC by filtering subjects whose labels did not completely match the neuroimaging measures.

Furthermore, significant FNC features can effectively reveal important interactions between functional networks. Therefore, we display the mean strength of the top 10 shared or unique significant FNC features ($p < 0.01$ with Bonferroni correction) in typical groups relative to original groups in the three source datasets, as shown in Figure 4. The specific mean strength, p values, and T-values of the above 20 significant FNC features using typical groups and original groups were outlined in Tables S5 and S6. Regarding the top 10 shared significant FNC features, the typical SZ (relative to typical HC) shows a greater decrease in five cerebellum-thalamus/caudate functional connectivities; a larger increase in four thalamus-related functional connectivities to superior temporal gyrus, left/right postcentral gyrus, and right middle occipital gyrus; and a larger increase in caudate-superior temporal gyrus functional connectivity than the original groups. More importantly, new findings in brain function were explored from typical groups relative to the original groups. For the top 10 unique significant FNC features within typical groups, typical SZ shows weakened connectivity in eight postcentral gyrus-related functional connectivities to the middle/superior temporal gyrus and inferior/right middle occipital gyrus and enhanced connectivity in two cerebellum-related functional connectivities to paracentral lobule and superior parietal lobule (compared with typical HC). The findings confirm that the typical groups can capture more significant, consistent, and novel between-group differences than the original groups across multiple datasets.

Taking another dataset (FBIRN, MPRC, or COBRE) as the independent dataset, the results (Figures S7–S12; Tables S7–S12) also manifested that the identified typical groups can capture more significant and consistent SZ-related functional connectivity abnormalities than the original groups. In summary, we believe that utilizing neuroimaging data with the assistance of clinical diagnosis to discover reliable typical subjects would contribute to exploring stable brain abnormalities and constructing reliable prediction models from multiple datasets. Thus, these separable typical subjects were used to construct a dimensional prediction model for unseen subjects.

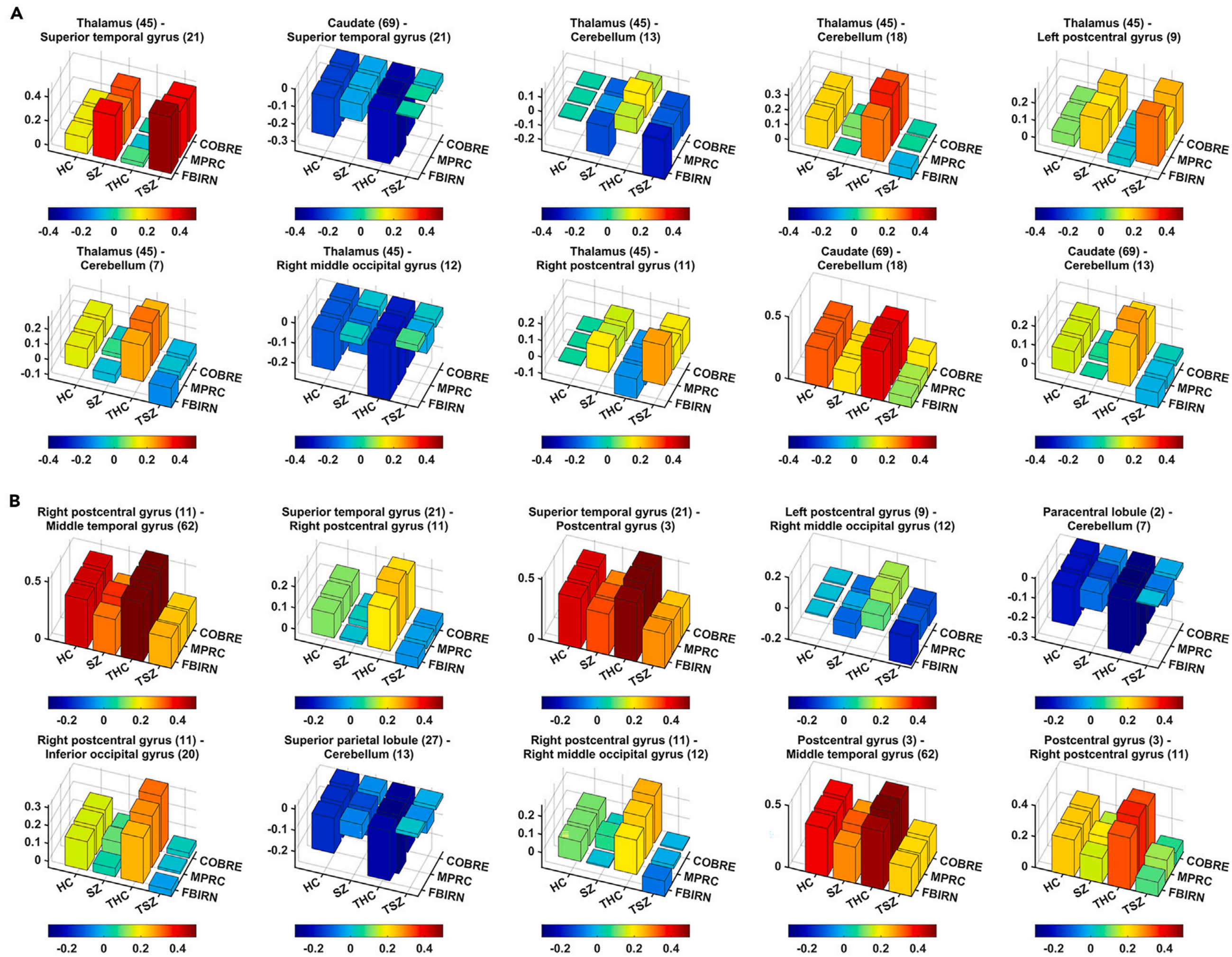


Figure 4. Mean strength of FNC features with significant inter-group differences for original groups and typical groups in the source datasets, including FBIRN, MPRC, and COBRE datasets

(A) Mean strength of the top 10 shared significant FNC features within original groups and typical groups across the source datasets.

(B) Mean strength of the top 10 unique significant FNC features within typical groups across the source datasets. The figure in parentheses represents the corresponding functional network ID. THC and TSZ represent typical HC and SZ groups, respectively. The typical groups (THC vs. TSZ) show more noticeable differences in the mean strength of the 20 FNC features than the original groups (HC vs. SZ).

The dimensional prediction model performs well in predicting independent subjects

We constructed a dimensional prediction model based on those valid typical subjects who presented remarkable and consistent inter-group differences across datasets to provide reliable dimensional scores indicating changes in brain function for independent subjects (see Figures 1D and S1). Specifically, we provided each independent subject with a stable comprehensive score by averaging multiple separate scores derived from its similarity to typical HC and typical SZ groups in each source dataset. It is worth noting that the comprehensive score and the three separate scores for each independent subject range from −1 to 1, indicating the absence of abnormality to the presence of significant abnormality. The stability of scores for independent subjects was adequately evaluated to demonstrate the feasibility of the dimensional prediction model, as shown in Figure 1F. Results indicate the reliability of dimensional scores derived from the model in revealing changes in brain function among independent subjects. Specifically, regarding the scores of independent subjects in BSNIP dataset, a strong correlation (mean correlation coefficient $r = 0.96$) is presented by calculating the Pearson correlation coefficients among their separate scores and comprehensive scores (Figure 5A). Figures S13–S15 show that, when using FBIRN, MPRC, and COBRE as the independent dataset, the mean correlation coefficient among the four sets of scores was 0.97, 0.93, and 0.96, respectively. The strong correlation among the scores of each independent subject highlights the within-group homogeneity and between-group heterogeneity of typical subjects identified from various datasets, affirming the stability and reliability of the dimensional prediction model.

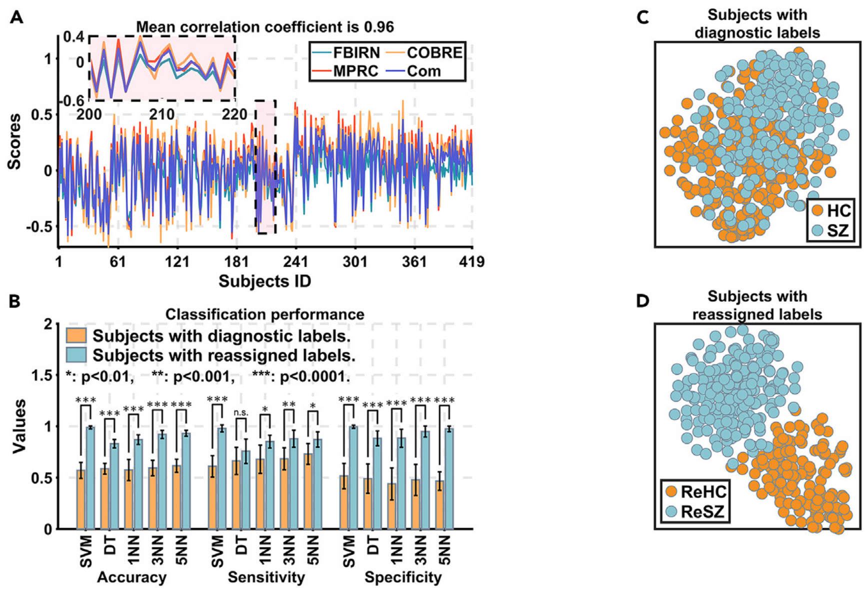


Figure 5. The validation of the dimensional prediction model using the independent BSNIP dataset

(A) Four sets of predicted continuous scores, including three sets of separate scores guided by the typical subjects in each source dataset and a set of comprehensive scores (Com) obtained based on the separate scores, for the independent subjects in the BSNIP dataset. The horizontal axis represents each independent subject, and the vertical axis represents the corresponding prediction score value of the subject. A portion of the diagram is enlarged in the upper left corner for better visualization. A strong correlation is shown between the four sets of scores.

(B) Average classification accuracy, sensitivity, and specificity for the original groups with diagnostic labels and relabeled groups with reassigned labels in the independent dataset based on the five classifiers, including SVM, DT, 1NN, 3NN, and 5NN classifiers. Asterisks (*) above the bars denote a statistically significant difference (i.e., $p < 0.01$) via two-sample t-tests in the classification performance of relabeled groups relative to original groups, with more asterisks being more significant. And n.s. above the bars represents no statistically significant differences (i.e., $p \geq 0.01$) in their classification performance. The classification performance using relabeled groups is significantly better than that using the original groups.

(C) 2D projection of the independent subjects with the diagnostic labels.

(D) 2D projection of the independent subjects with the reassigned labels. ReHC and ReSZ represent the relabeled HC group and relabeled SZ group in the independent dataset, respectively. In (C and D), the relabeled groups (ReHC vs. ReSZ) are more separable compared with the original groups (HC vs. SZ).

We relabeled the independent subjects to validate that the proposed model can also result in distinguishable groups, which holds practical significance for the treatment and prognosis of mental disorders. Based on the comprehensive score, each independent subject was categorized into the relabeled HC group, the relabeled SZ group, and the Boundary group in which subjects were middle and could not be confidently categorized into a specific group. Using BSNIP as the independent dataset for illustration, intra-group compactness and inter-group separability of the relabeled HC and relabeled SZ groups were evaluated from the following aspects. Under an unbiased 10-fold cross-validation classification framework, the relabeled groups show significantly improved average classification performances over the original groups based on the five classifiers (see Figure 5B). More classification evaluations were shown in Tables S13–S17. Similarly, we employed t-SNE to map subjects with the diagnostic labels (Figure 5C) and subjects with reassigned labels (Figure 5D) into a 2D projection to reveal the natural structure in the independent dataset. The results indicated that relabeled subjects in the same group were closely clustered together and could be distinguished from the other group, while there was considerable overlap among the diagnosis-labeled subjects in different groups. We also demonstrated that relabeled groups achieved enhanced intra-group compactness and inter-group separability relative to original groups by comparing five widely used evaluation metrics (Table S18). In short, the relabeled groups show improved intra-group compactness and inter-group separability compared with the original groups.

Using another dataset (FBIRN, MPRC, or COBRE) as the independent dataset, the results (Figures S13–S15; Table S18) also support that the proposed dimensional prediction model was promising in characterizing brain function changes. In conclusion, guided by the diagnostic information of the typical subjects, our method provides dimensional scores that reveal the degree of pathology from the neuroimaging perspective.

Relabeled independent subjects show more distinct inter-group differences by jointly using neuroimaging measures

To validate that the proposed dimensional prediction model can also result in groups with greater inter-group differences from a neuroimaging perspective, we explored FNC features that significantly differed between relabeled HC and relabeled SZ in the independent dataset using two-sample t-tests. Due to limited space, we only display the significant FNC features identified in the independent BSNIP dataset for illustration. T-values of significant FNC features ($p < 0.01$ with or without Bonferroni correction) between relabeled groups (Figure 6B)

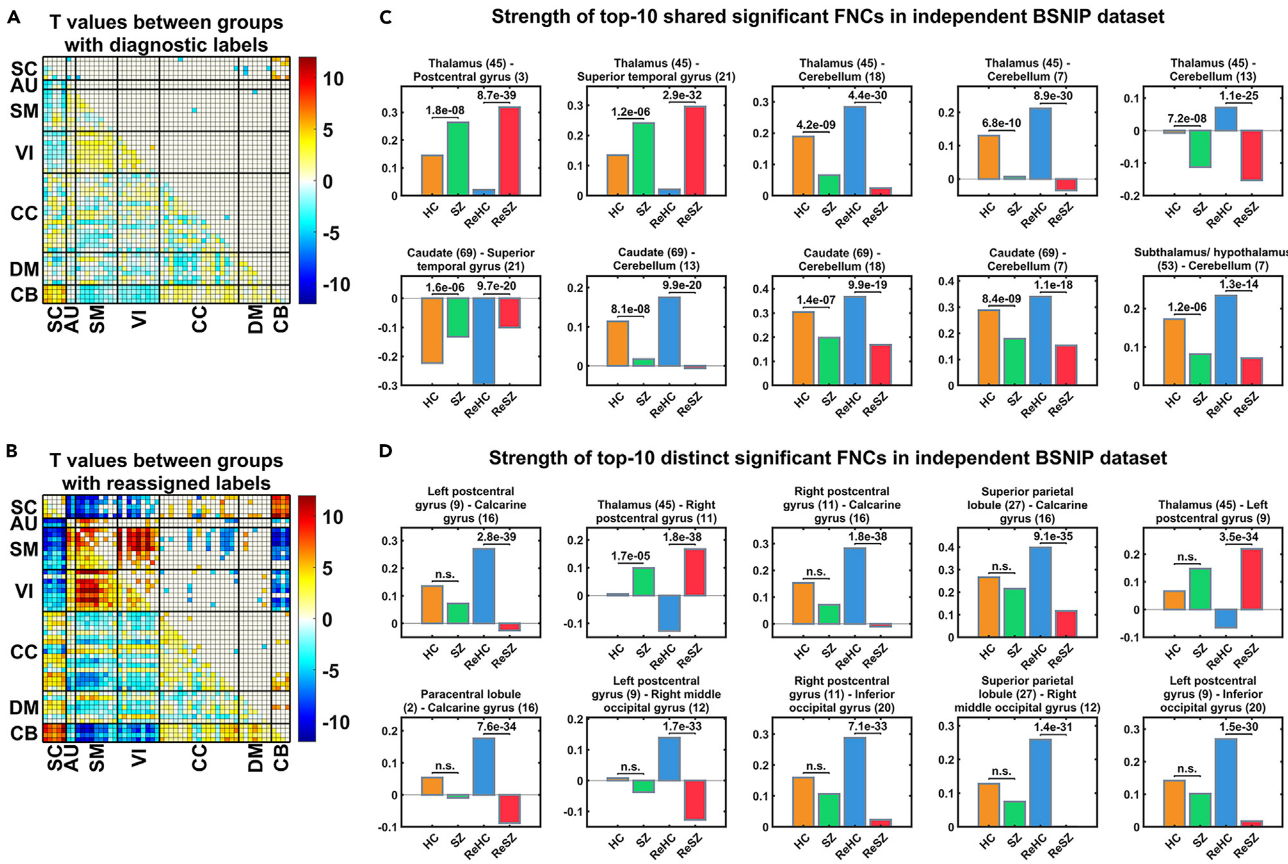


Figure 6. Inter-group differences within the independent BSNIP dataset

(A and B) Two-sample t-test T-values for the significant FNC features (upper triangle: $p < 0.01$ with Bonferroni correction, lower triangle: no correction) within original groups and within relabeled groups, respectively. In (A and B), 53 intrinsic connectivity networks are divided into seven brain functional domains, including sub-cortical (SC), auditory (AU), sensorimotor (SM), visual (VI), cognitive-control (CC), default-mode (DM), and cerebellar (CB) domains. Regarding (A and B), the inter-group differences in the relabeled groups are more noticeable than that in the original groups.

(C) Mean strength of the top 10 shared significant FNC features of subjects in original groups and in relabeled groups.

(D) Mean strength of the top 10 unique significant FNC features of subjects in relabeled groups relative to original groups. The numbers above the bars in (C and D) are two-sample t-test p values between groups, and n.s. above the bars represents no significant differences ($p \geq 0.01$ with Bonferroni correction) between the two groups. The figure in parentheses represents the corresponding functional network ID. ReHC and ReSZ represent the relabeled HC group and relabeled SZ group in the independent dataset, respectively. In (C and D), the differences in mean strength of the 20 FNC features are more noticeable in the relabeled groups (ReHC vs. ReSZ) compared with that in the original groups (HC vs. SZ).

suggest a more prominent group difference than those between the original groups (Figure 6A). More visually, Figures 6C and 6D display the mean strength of the top 10 shared or unique significant FNC features within relabeled groups relative to the original groups in the independent dataset. Meanwhile, Tables S19 and S20 outlined the specific mean strength, p values, and T-values of the above 20 significant FNC features using original groups and relabeled groups. These results demonstrated more evident differences between the relabeled groups relative to the original groups. Specifically, compared with the top 10 shared significant FNC features within the original groups, relabeled SZ (relative to relabeled HC) shows a larger decrease in cerebellum-thalamus/caudate functional connectivity and a larger increase in thalamus-superior temporal gyrus/postcentral gyrus functional connectivity. Similar findings were derived from the typical groups and original groups in the source datasets (Figure 4A). Additionally, for the top 10 unique significant FNC features within the relabeled groups, relabeled SZ shows connectivity differences in the parietal lobe (i.e., superior parietal lobule, paracentral lobule, and postcentral gyrus) and occipital lobe (i.e., right middle occipital gyrus and inferior occipital gyrus) compared with the relabeled HC. These findings were consistent with the unique significant FNC features within the typical groups in source datasets (Figure 4B). Additionally, significantly weaker connectivity connecting the calcarine gyrus to the right/left postcentral gyrus, paracentral lobule, and superior parietal lobule was observed in relabeled SZ (relative to relabeled HC), which was insignificant in the original groups. These findings suggest that the relabeled groups can capture more significant and novel FNC features revealing impaired brain function in SZ than the original groups.

Using other datasets (FBIRN, MPRC, or COBRE) as the independent dataset, differences between relabeled groups were more prominent relative to original groups (Figures S16–S18; Tables S21–S26). Therefore, there might be inconsistency between diagnostic labels and

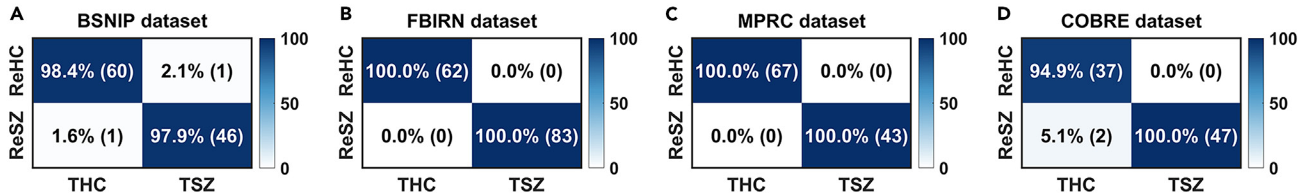


Figure 7. Confusion matrices between the labels of typical groups and the relabeled groups in the same dataset
(A–D) The corresponding confusion matrices for BSNIP, FBIRN, MPRC, and COBRE, respectively. THC and TSZ represent typical HC and typical SZ groups, respectively. ReHC and ReSZ represent relabeled HC and relabeled SZ groups, respectively.

neuroimaging measures within the independent datasets, resulting in inconspicuous differences between the original groups. In short, both typical groups and relabeled groups can reveal more stable, consistent, and significant brain functional impairments, supporting that the combination of diagnostic information and neuroimaging measures in our method enables more accurate predictions.

More reliable biomarkers are found for SZ

Since the present work suggested that both typical subjects in the source datasets and relabeled subjects in the independent dataset had matching labels and neuroimaging measures, we aimed to verify the consistency and stability of our proposed method by constructing a confusion matrix between the labels of these subjects. As shown in Figure 1, we employed the leave-one-dataset-out cross-validation procedure to iteratively assess the performance of our LAMP method. Taking the BSNIP dataset as an example, in the first iteration, it served as the independent dataset to yield relabeled subjects, while, in the subsequent iteration, it was used as the source dataset to identify typical subjects within it. We computed a confusion matrix between the labels of typical subjects and relabeled subjects in BSNIP dataset to verify the consistency between these subjects. As shown in Figure 7, the labels of the majority of typical subjects in each dataset remained unchanged after relabeling. These findings confirm the consistency between typical subjects and relabeled subjects, and support the stability and reliability of our proposed method.

To identify the putative biomarkers revealing abnormal brain function in SZ patients compared to HCs, we investigated the significant and consistent inter-group differences in FNC using typical subjects and relabeled subjects across the four datasets. Two-sample t-tests were utilized to calculate inter-group differences in FNC features for original groups, typical groups, and relabeled groups in each dataset. We analyzed and displayed the mean strength, p values, and T-values of the top 8 FNC features that showed significant and shared inter-group differences across the original groups, typical groups, and relabeled groups in all four datasets (see Figure 8A; Table S27). Here, we display the top 8 significant FNC features since only eight were shared among original groups, typical groups, and relabeled groups across the four datasets. Our findings indicate that SZ group consistently exhibited weaker connectivity between the cerebellum and thalamus/caudate, as well as increased connectivity between the superior temporal gyrus and thalamus/caudate (compared with HC group). Additionally, we also analyzed and exhibited the mean strength, p values, and T-values of the top 10 FNC features that exhibited significant inter-group differences in the typical groups and relabeled groups across the four datasets but were not significant in the original groups of any dataset (see Figure 8B; Table S28). Compared with the original group, both typical SZ and relabeled SZ groups showed weaker connectivity related to the parietal lobe and occipital lobe, including right/left postcentral gyrus-inferior/right middle occipital gyrus connectivity, superior parietal lobule-right middle occipital gyrus connectivity, and paracentral lobule-inferior/right middle occipital gyrus connectivity, and significantly stronger paracentral lobule-cerebellum connectivity than typical HC and relabeled HC groups. In sum, by utilizing typical subjects and relabeled subjects whose labels are more aligned with the neuroimaging measures, our method enables us to find significant, consistent, and reliable inter-group differences across various datasets. These differences hold promise as putative biomarkers indicating aberrant brain functional connectivity associated with SZ.

Reproducible results on SZ-HC data are observed through the replication experiment

To further validate the reproducibility of the proposed LAMP method, we conducted a replication experiment using the same four datasets and the identical experimental workflow, as depicted in Figure 1. We evaluated the reproducibility of the proposed method by comparing the consistency between the results from the two separate runs. Specifically, we compared the consistency of typical subjects identified in each source dataset between the two runs, the consistency of relabeled subjects in each independent dataset between the two runs, and the consistency of explored biomarkers between the two runs. The experimental results exhibited slight variation due to the randomness introduced by constructing multiple complete random decision trees in the CRF-based model.

In the replication experiment, we first observed a significant overlap between the typical subjects identified from each source dataset in the two separate runs, as shown in Figure S19. It is noteworthy that none of the subjects identified as typical HC (or SZ) in the first run were identified as typical SZ (or HC) in the second run for each source dataset. This suggests that the identified typical subjects were stable and reproducible. Additionally, the reproducibility of the dimensional prediction model was validated by assessing the consistency of results between two runs for relabeled subjects in each independent dataset, as shown in Figure 9. Taking the results corresponding to the independent BSNIP dataset as an example, among the 146 relabeled HCs (ReHC-1) in the first run, 141 subjects were also relabeled as HCs (ReHC-2) in

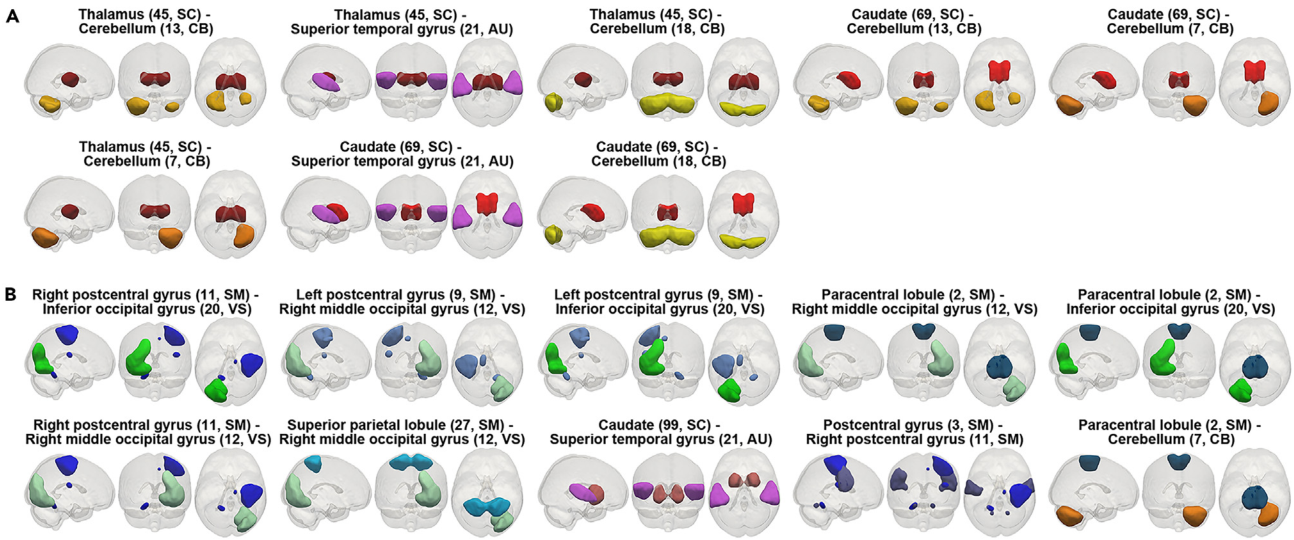


Figure 8. The functional networks with significant inter-group differences in connectivity across the four datasets in the brain

(A) Top 8 shared significant FNC features within subjects in the original groups, typical groups, and relabeled groups across the four datasets.

(B) Top 10 unique significant FNC features within subjects in typical groups and relabeled groups across the four datasets compared with the original groups. 53 intrinsic connectivity networks are divided into seven brain functional domains, including sub-cortical (SC), auditory (AU), sensorimotor (SM), visual (VI), cognitive-control (CC), default-mode (DM), and cerebellar (CB) domains. The figure in parentheses represents the corresponding functional network ID.

the second run. Additionally, out of the 195 relabeled SZ subjects (ReSZ-1) in the first run, 192 subjects were also relabeled as SZ subjects (ReSZ-2) in the second run. Although a few subjects classified into the Boundary group showed slight variations between the two runs, no subjects that were relabeled as HC (or SZ) in the first run were relabeled as SZ (or HC) in the second run. This indicates the reproducibility of the dimensional prediction model, leading to stable outcomes. Furthermore, we also evaluated the consistency of the explored biomarkers in the two separate runs to validate the robustness and reproducibility of putative biomarkers. In both runs, the LAMP method explored 18 putative biomarkers: (1) 8 FNC features that exhibited significant inter-group differences between the original groups, between the typical groups, and between the relabeled groups in all four datasets and (2) 10 FNC features that displayed significant inter-group differences between the typical groups and between the relabeled groups across the four datasets, while not showing significant inter-group differences between the original groups in any dataset. As shown in Figure S20 and Table S29, 17 out of the 18 biomarkers were consistently detected in both runs. Among the 17 consistently detected biomarkers, SZ group (compared to HC group) primarily demonstrated abnormal connectivity related to the cerebellum, thalamus, caudate, parietal lobe, and occipital lobe. For the one biomarker that was not consistently identified, SZ group exhibited weakened connectivity between the paracentral lobule and inferior occipital gyrus in the first run and enhanced connectivity between the caudate and left postcentral gyrus in the second run, compared to the HC group. In conclusion, the reproducible experimental results effectively demonstrated the reproducibility and stability of the proposed method.

The generalizability of the proposed LAMP method applied to autism spectrum disorder (ASD) and HC data is verified

To validate the generalizability of our LAMP method for other disorders, we applied our LAMP method to fMRI data of subjects with ASD and age-matched HCs using the same experimental workflow (see Figure 1). More specifically, four datasets named SubData1, SubData2, SubData3, and SubData4 were used from the Autism Brain Imaging Data Exchange (ABIDEI) data. We extracted FNC features for each subject via NeuroMark and carefully removed nuisance effects, including age, gender, motion, and site effects (see Table S30 for demographic information in detail).

Just like the evaluation on SZ-HC data, we investigated inter-group differences, classification performance, between-group separation, within-group cohesion, and data visualization for typical ASD and typical HC groups identified from each source dataset, as shown in Tables S31 and S32 and Figures S21–S27. Our findings indicated enhanced intra-group compactness and inter-group separability, as well as significant and consistent inter-group differences between the typical ASD and typical HC groups compared to the original ASD and HC groups in each source dataset. Substantial experiments demonstrated the separable data structure of typical subjects in each source dataset, reflecting clear consistency between their labels and fMRI measures. Furthermore, a dimensional prediction model for ASD-HC data was constructed to provide reliable dimensional scores indicating changes in brain function for independent subjects. Taking the subjects in the independent SubData2 dataset for illustration, a strong correlation (mean correlation coefficient $r = 0.87$) is presented among their separate scores and comprehensive scores in Figure 10A. Similarly, we obtained relabeled ASD group and relabeled HC group in the independent dataset based on the comprehensive scores and evaluated their separability through multiple experiments. As shown in Figure 10B, using an unbiased 10-fold cross-validation, the classification performance on the relabeled groups outperformed that on the original groups. In

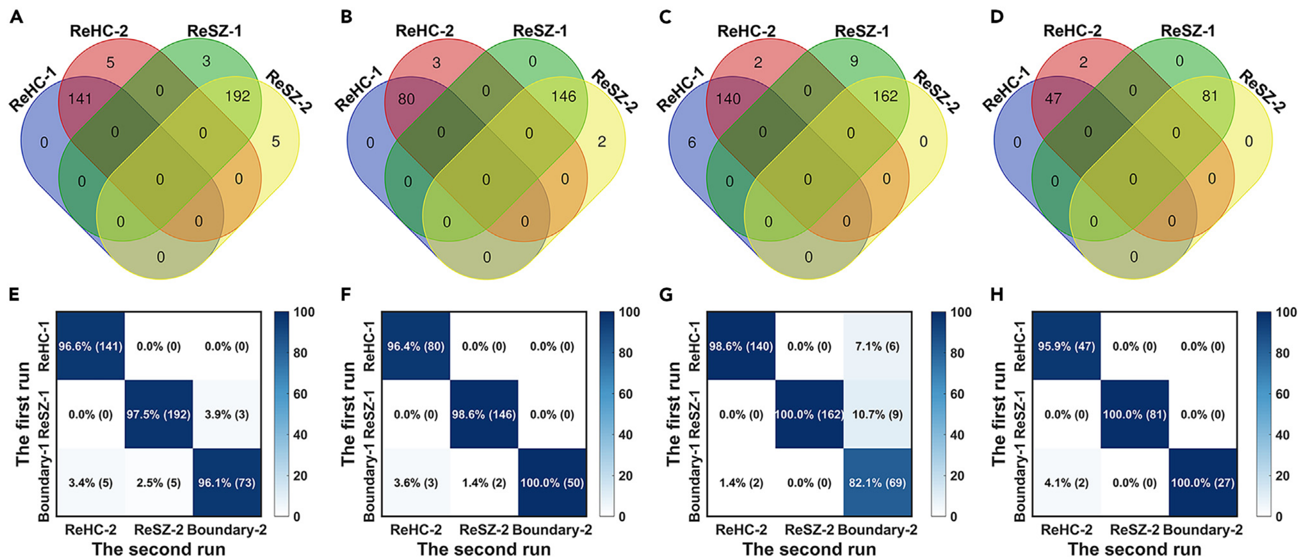


Figure 9. The overlap of the relabeled subjects in each independent dataset between the two separate runs

(A–D) The Venn diagrams display the number of relabeled subjects (i.e., relabeled HC and relabeled SZ subjects) that overlap between the two runs in BSNIP, FBIRN, MPRC, and COBRE, respectively.

(E–H) The confusion matrices display the number of relabeled subjects (i.e., subjects in relabeled HC, relabeled SZ, and Boundary groups) that overlap between the two runs in BSNIP, FBIRN, MPRC, and COBRE, respectively. The numbers in parentheses represent the number of subjects in each respective category. ReHC-1, ReSZ-1, and Boundary-1 represent the relabeled HC subjects, relabeled SZ subjects, and subjects in the Boundary group identified in each independent dataset during the first run. Similarly, ReHC-2, ReSZ-2, and Boundary-2 represent the relabeled HC subjects, relabeled SZ subjects, and subjects in the Boundary group identified in each independent dataset during the second run.

addition, by utilizing t-SNE visualization technology, it can be observed that relabeled subjects within the same group clustered closely together and could be easily differentiated from subjects in another group (Figure 10D). Conversely, a significant overlap was observed among the diagnosis-labeled subjects from different groups (Figure 10C). In short, compared to the original ASD and HC groups, the improved intra-group compactness, inter-group separability, and inter-group differences between the relabeled ASD and relabeled HC groups were verified for each independent dataset. The findings were shown in Table S33 and Figures S28–S34, which demonstrated the stability and reliability of the dimensional prediction model on ASD-HC data.

More importantly, after confirming the consistency of labels between typical subjects and relabeled subjects from the same dataset (Figure S35), we utilized these subjects with matching labels and neuroimaging measures to explore biomarkers for ASD. As outlined in Table S34, compared to the HC group across the four datasets, ASD group showed weakened connectivity between the caudate and cerebellum, thalamus and cerebellum, postcentral gyrus and left postcentral gyrus, and postcentral gyrus and superior parietal lobule. In addition, ASD group showed increased connectivity between the thalamus and postcentral gyrus, caudate and postcentral gyrus, and caudate and superior temporal gyrus. In sum, by utilizing typical subjects and relabeled subjects whose labels were more aligned with the neuroimaging measures, our method enables us to find significant, consistent, and reliable inter-group differences across various datasets. These differences hold promise as putative biomarkers indicating aberrant brain functional connectivity associated with ASD.

In conclusion, the experimental results (Tables S31–S34; Figures S21–S35) support the successful extension of our method to other disorders. By utilizing the proposed method, we explored significant, consistent, and stable biomarkers for ASD and provided dimensional scores indicating abnormal brain function associated with ASD. This demonstrates the robustness and generalizability of our method on different populations.

DISCUSSION

Potential inconsistency between diagnostic labels and neuroimaging measures negatively impacts neuroscience-based research of exploring reliable biomarkers and constructing accurate prediction models for mental disorders.[3,29] In this paper, we proposed a neuroimaging-based LAMP method to characterize brain change and explore putative biomarkers for mental disorders only using subjects whose labels are consistent with measures from multiple datasets. It helps build a discriminable prediction model using the identified typical subjects with clear separability as benchmarks. Improved intra- and inter-group structures of the typical groups confirmed by substantial experimental results indicate greater alignment between labels and fMRI measures, potentially enabling the construction of a reliable prediction model for SZ. Impressively, compared with the original groups, the average classification accuracy on the typical groups within and across source datasets increased by 52.08% and 47.03% (Tables S35–S39), respectively, demonstrating the outstanding separability and generalizability of the

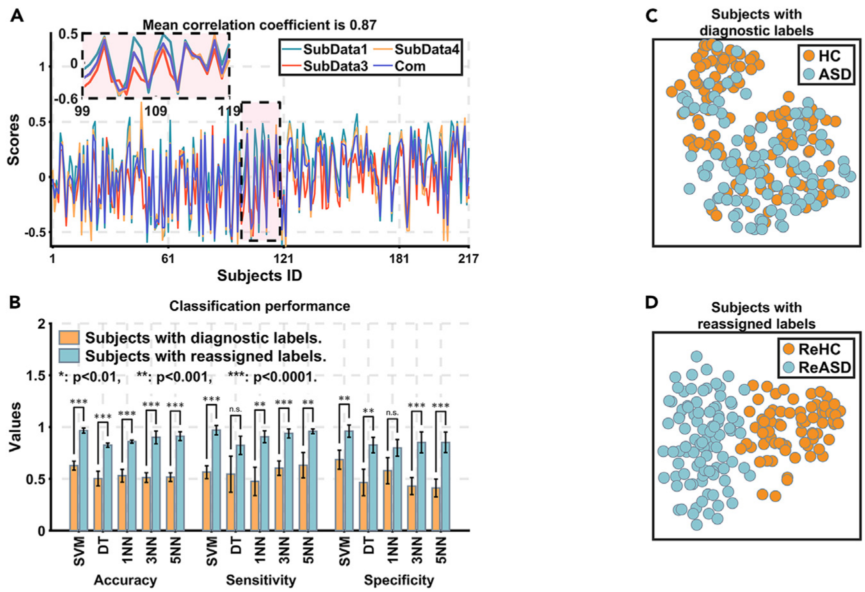


Figure 10. The validation of the dimensional prediction model using the independent SubData2 dataset including ASD and HC subjects

(A) Four sets of predicted continuous scores, including three sets of separate scores guided by the typical subjects in each source dataset and a set of comprehensive scores (Com) obtained based on the separate scores, for the independent subjects in the SubData2 dataset. The horizontal axis represents each independent subject, and the vertical axis represents the corresponding prediction score value of the subject. A portion of the diagram is enlarged in the upper left corner for better visualization. A strong correlation is shown between the four sets of scores.

(B) Average classification accuracy, sensitivity, and specificity for the original groups with diagnostic labels and relabeled groups with reassigned labels in the independent dataset based on the five classifiers, including SVM, DT, 1NN, 3NN, and 5NN classifiers. Asterisks (*) above the bars denote a statistically significant difference (i.e., $p < 0.01$) via two-sample t-tests in the classification performance of relabeled groups relative to original groups, with more asterisks being more significant. And n.s. above the bars represents no statistically significant differences (i.e., $p \geq 0.01$) in their classification performance. The classification performance using relabeled groups is significantly better than that using the original groups.

(C) 2D projection of the independent subjects with the diagnostic labels.

(D) 2D projection of the independent subjects with the reassigned labels. ReHC and ReASD represent the relabeled HC group and relabeled ASD group in the independent dataset, respectively. In (C and D), the relabeled groups (ReHC vs. ReASD) are more separable compared with the original groups (HC vs. ASD).

preserved subjects. More importantly, the proposed dimensional prediction model offers a comprehensive scoring system for each independent subject, quantifying the extent of brain abnormalities from negligible to severe. Based on the predicted scores, we categorized independent subjects into patient group, normal group, and Boundary group with minor brain alterations, which is meaningful for understanding the underlying pathogenesis of disorders and can also aid in the accurate diagnosis for mental disorders. Numerous experiments demonstrated the effectiveness and robustness of the model. For each independent dataset, the average correlation among the predicted scores derived from typical groups in different source datasets was over 0.9, demonstrating the homogeneity of typical groups and the stability of the model. In addition, considerable overlap between the labels of subjects in typical groups and relabeled groups from the same dataset manifested the consistency of outcomes and reliability of our method. It is worth noting that some HCs showed abnormalities in brain function and were relabeled as SZ or placed in the Boundary group, which supports existing findings of the possible disease risk in HCs.[2,29] We also demonstrated significantly improved inter-group separability, intra-group compactness, and average classification accuracy (increased by about 50%) of relabeled HC and SZ groups relative to diagnosis-labeled groups in various independent datasets. It suggested that the reassigned labels quite agreed with the neurobiological substrates within independent datasets from a neuroimaging perspective, avoiding the degeneration of separability caused by the inconsistency. The replication experiment on SZ-HC data and the generalizability validation experiment on ASD-HC data further demonstrated the feasibility of the LAMP method. In brief, the proposed LAMP method provided a unitary, accurate, and dimensional approach to assess the pathology of mental disorders, which might better conform to clinical reality and has the potential to assist traditional categorical diagnostic methods.

In addition, utilizing inter-group differences derived from subjects with matching labels and measures from multiple datasets helps uncover more reliable, consistent, and stable potential biomarkers indicating abnormal brain functional connectivity associated with the disorder. We explored significant, reliable, and consistent brain functional impairments in SZ patients as putative biomarkers across multiple datasets. Specifically, weaker connectivity between sub-cortical (thalamus and caudate) and cerebellar domains and increased connectivity between sub-cortical (thalamus and caudate) and auditory (superior temporal gyrus) domains were consistently found in original SZ, typical SZ, and relabeled SZ groups relative to HCs across the four datasets, which can be regarded as stable biomarkers to discriminate SZ patients from HCs. Indeed, thalamus, caudate, and cerebellar are linked to brain information processing, guiding behavior, and cognition, respectively, showing significant abnormalities in SZ patients.[41–43] Previous studies have reported similar functional abnormalities in SZ patients but lacked adequate generalizability verification across datasets, unfortunately.[44,45] More importantly, the typical SZ and relabeled SZ groups across the four datasets displayed marked functional alterations that were insignificant in the original SZ group relative to HCs. This indicates that more significant inter-group differences observed in both the identified typical groups and relabeled groups based on our method helped to explore new reliable abnormalities in SZ patients. Particularly, hypo-connectivity between the parietal lobe (i.e., postcentral gyrus, superior parietal lobule, and paracentral lobule) in the sensorimotor domain and right middle/inferior occipital gyrus in the visual domain was detected in typical SZ and relabeled SZ groups. Parietal lobe and occipital cortex have key roles in maintaining visuospatial information; disturbed parieto-occipital functional connectivity is related to positive symptoms, such as cognitive deficits, disorganization, and delusions, in SZ patients.[46–48] In addition, hyper-connectivity between the paracentral lobule and cerebellum was found in typical SZ and relabeled SZ groups relative to HCs. This finding suggests that the functional association between the cerebellar and sensorimotor domains is affected in SZ, which may be related to disorganized or abnormal motor behavior. In short, we explored more new, significant, and consistent brain functional impairments in both typical SZ and relabeled SZ groups relative to original groups based on our LAMP method, promoting the investigation of neuropathological substrates on mental disorders.

Furthermore, it is important to note that clinical diagnostic labels may not directly align with the neuroimaging measures. As a result, there can be varying inter-group differences within the original groups across different datasets, making it challenging to consider the differences as reliable biomarkers for clinical use. To address this issue, we leveraged typical subjects and relabeled subjects who had matching labels and fMRI measures to enhance the inter-group differences, thus enabling us to identify biomarkers that consistently exhibited significant inter-group differences across multiple datasets. Particularly, ASD and SZ are two brain disorders that share considerable clinical and neuroimaging features, making it difficult to distinguish them from each other.[49] Using typical subjects in the source datasets and relabeled subjects in the independent dataset, we were able to highlight the inter-group differences between SZ and HC groups, as well as ASD and HC groups. Consequently, the differences between SZ and ASD groups were further amplified. As a result, we discovered unique biomarkers for ASD, such as weakened connectivity between the postcentral gyrus and superior parietal lobule compared to the HC group. In brief, more significant and consistent inter-group differences observed across datasets can serve as potential biomarkers to distinguish between different groups, aiding in indicating abnormal brain function associated with the disorders and assisting doctors in accurate diagnosis based on these biomarkers.

In summary, we propose a label-noise filtering-based dimensional prediction method, LAMP, which mitigates the impact of potential inconsistency between diagnostic labels and neuroimaging measures in exploring reliable biomarkers and constructing accurate prediction models. Substantial evidence highlights that LAMP method can explore stable and reliable functional abnormalities unveiling pathogenesis and build the dimensional prediction model revealing the degree of abnormalities for mental disorders using neuroimaging data. It is important to point out that, as more source datasets from various sites are incorporated, a greater number of reliable typical subjects are obtained from these source datasets. Consequently, the conclusions drawn from our LAMP method would become progressively more reliable, which partially mitigates the issue of limited typical subjects identified from a single source dataset. The LAMP method holds promise in understanding the underlying pathogenesis and assisting accurate diagnosis of mental disorders.

Limitations of the study

This study has several limitations. First, the proposed LAMP method has only been validated using fMRI data. In future work, more datasets including other functional and structural measures could be involved to comprehensively reflect the effects of mental disorders on brain function and structure. Besides, completely mitigating the nuisance effects is challenging as it is difficult to determine if these effects have been completely eliminated without any ground truth about group differences. This highlights the need for the development of more advanced algorithms. In addition, this work utilized neuroimaging measures to score independent subjects without incorporating symptom-based scores. Future research may benefit from integrating diagnostic scores and neuroimaging measures to generate more comprehensive dimensional scores. Finally, although sufficient experimental validation has demonstrated that integrating neuroimaging measures with clinical diagnostic information can help uncover more reliable findings, further validation of the proposed method is still needed in clinical practice.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Participants and preprocessing
  - Estimation of functional network connectivity (FNC) features from fMRI data
  - Overall workflow of the present work

- ○ The proposed label-noise filtering-based dimensional prediction (LAMP) method
- ○ Evaluation of the validity of the identified typical subjects in each source dataset
- ○ Evaluation of the feasibility of the dimensional prediction model for the independent dataset
- ○ Comparison of the typical subjects and the relabeled groups derived from the same dataset
- ○ Reproducibility evaluation on SZ-HC data through the replication experiment
- ○ Within and between datasets separability of the original groups and typical groups across SZ-HC datasets
- ○ The generalizability validation of the proposed LAMP method on ASD-HC data
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109319.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant no. 62076157 and 61703253 to Y.D.), Fund Program for the Scientific Activities of Selected Returned Overseas Professionals in Shanxi Province (grant no. 20210033 to Y.D.), the 1331 Engineering Project of Shanxi Province of China, and the National Institutes of Health (grant no. R01MH118695, R01MH123610, and NSF 2112455).

AUTHOR CONTRIBUTIONS

Conceptualization, Y.X. and Y.D.; methodology, Y.X. and Y.D.; software, Y.X.; formal analysis, Y.X. and Y.D.; investigation, Y.X. and Y.D.; data curation, Y.D., P.K., and T.G.M.v.E.; writing – original draft, Y.X. and Y.D.; writing – review & editing, Y.X., Y.D., T.G.M.v.E., P.K., G.D.P., and V.D.C.; supervision, Y.D.; funding acquisition, Y.D.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 29, 2023
Revised: September 17, 2023
Accepted: February 19, 2024
Published: February 23, 2024

REFERENCES

1. GBD 2019 Mental Disorders Collaborators (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatr. *9*, 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3.
2. Rashid, B., and Calhoun, V. (2020). Towards a brain-based predictome of mental illness. Hum. Brain Mapp. *41*, 3468–3535. https://doi.org/10.1002/hbm.25013.
3. Cuthbert, B.N., and Insel, T.R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. *11*, 126. https://doi.org/10.1186/1741-7015-11-126.
4. Craddock, N., and Mynors-Wallis, L. (2014). Psychiatric diagnosis: impersonal, imperfect and important. Br. J. Psychiatry *204*, 93–95. https://doi.org/10.1192/bjp.bp.113.133090.
5. Frénay, B., and Verleysen, M. (2014). Classification in the presence of label noise: a survey. IEEE Transact. Neural Networks Learn. Syst. *25*, 845–869. https://doi.org/10.1109/Tnnls.2013.2292894.
6. Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Zhou, Q., Liu, T., and Ge, Z. (2022). Improving medical images classification with label noise using dual-uncertainty estimation. IEEE Trans. Med. Imaging *41*, 1533–1546. https://doi.org/10.1109/Tmi.2022.3141425.
7. Schnack, H.G., and Kahn, R.S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. Front. Psychiatry *7*, 50. https://doi.org/10.3389/fpsyt.2016.00050.
8. Shatte, A.B.R., Hutchinson, D.M., and Teague, S.J. (2019). Machine learning in mental health: a scoping review of methods and applications. Psychol. Med. *49*, 1426–1448. https://doi.org/10.1017/S0033291719000151.
9. Nielsen, A.N., Barch, D.M., Petersen, S.E., Schlaggar, B.L., and Greene, D.J. (2020). Machine learning with neuroimaging: evaluating its applications in psychiatry. Biol. Psychiatry. Cogn. Neurosci. Neuroimaging *5*, 791–798. https://doi.org/10.1016/j.bpsc.2019.11.007.
10. Li, F., Sun, H., Biswal, B.B., Sweeney, J.A., and Gong, Q. (2021). Artificial intelligence applications in psychoradiology. Psychoradiology *1*, 94–107. https://doi.org/10.1093/psyrad/kkac003.
11. Anderson, J.S., Nielsen, J.A., Froehlich, A.L., DuBray, M.B., Druzgal, T.J., Cariello, A.N., Cooperrider, J.R., Zielinski, B.A., Ravichandran, C., Fletcher, P.T., et al. (2011). Functional connectivity magnetic resonance imaging classification of autism. Brain *134*, 3742–3754. https://doi.org/10.1093/brain/awr263.
12. Cao, B., Cho, R.Y., Chen, D., Xiu, M., Wang, L., Soares, J.C., and Zhang, X.Y. (2020). Treatment response prediction and individualized identification of first-episode drug-naive schizophrenia using brain functional connectivity. Mol. Psychiatry *25*, 906–913. https://doi.org/10.1038/s41380-018-0106-5.
13. Cai, X.L., Xie, D.J., Madsen, K.H., Wang, Y.M., Bögemann, S.A., Cheung, E.F.C., Møller, A., and Chan, R.C.K. (2020). Generalizability of machine learning for classification of schizophrenia based on resting-state functional MRI data. Hum. Brain Mapp. *41*, 172–184. https://doi.org/10.1002/hbm.24797.
14. Rodrigue, A.L., Mastrovito, D., Esteban, O., Durnez, J., Koenis, M.M.G., Janssen, R., Alexander-Bloch, A., Knowles, E.M., Mathias, S.R., Mollon, J., et al. (2021). Searching for imaging biomarkers of psychotic dysconnectivity. Biol. Psychiatry. Cogn. Neurosci. Neuroimaging *6*, 1135–1144. https://doi.org/10.1016/j.bpsc.2020.12.002.
15. Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., and Koutsouleris, N. (2015). Detecting neuroimaging biomarkers for schizophrenia: A meta-analysis of multivariate pattern recognition studies. Neuropsychopharmacology *40*, 1742–1751. https://doi.org/10.1038/npp.2015.22.
16. Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., et al. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of

depression. Nat. Med. *23*, 28–38. https://doi.org/10.1038/nm.4246.
17. Grisanzio, K.A., Goldstein-Piekarski, A.N., Wang, M.Y., Rashed Ahmed, A.P., Samara, Z., and Williams, L.M. (2018). Transdiagnostic symptom clusters and associations with brain, behavior, and daily function in mood, anxiety, and trauma disorders. JAMA Psychiatr. *75*, 201–209. https://doi.org/10.1001/jamapsychiatry.2017.3951.
18. Chang, M., Womer, F.Y., Gong, X., Chen, X., Tang, L., Feng, R., Dong, S., Duan, J., Chen, Y., Zhang, R., et al. (2021). Identifying and validating subtypes within major psychiatric disorders based on frontal–posterior functional imbalance via deep learning. Mol. Psychiatry *26*, 2991–3002. https://doi.org/10.1038/s41380-020-00892-3.
19. Pelin, H., Ising, M., Stein, F., Meinert, S., Meller, T., Brosch, K., Winter, N.R., Krug, A., Leenings, R., Lemke, H., et al. (2021). Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning. Neuropsychopharmacology *46*, 1895–1905. https://doi.org/10.1038/s41386-021-01051-0.
20. Dinga, R., Schmaal, L., Penninx, B.W.J.H., van Tol, M.J., Veltman, D.J., van Velzen, L., Mennes, M., van der Wee, N.J.A., and Marquand, A.F. (2019). Evaluating the evidence for biotypes of depression: Methodological replication and extension of. Neuroimage. Clin. *22*, 101796. https://doi.org/10.1016/j.nicl.2019.101796.
21. Winter, N.R., and Hahn, T. (2022). Significance and stability of deep learning-based identification of subtypes within major psychiatric disorders. Mol. Psychiatry *27*, 1858–1859. https://doi.org/10.1038/s41380-022-01482-1.
22. Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A.M., Nigg, J.T., and Fair, D.A. (2019). The heterogeneity problem: approaches to identify psychiatric subtypes. Trends Cogn. Sci. *23*, 584–601. https://doi.org/10.1016/j.tics.2019.03.009.
23. Zhang, X., Wu, X., Chen, F., Zhao, L., and Lu, C.T. (2020). Self-paced robust learning for leveraging clean labels in noisy data. Proc. AAAI Conf. Artif. Intell. *34*, 6853–6860.
24. Wu, P., Zheng, S., Goswami, M., Metaxas, D., and Chen, C. (2020). A topological filter for learning with label noise. Adv. Neural Inf. Process. Syst. *33*, 21382–21393.
25. Wei, Q., Sun, H., Lu, X., and Yin, Y. (2022). Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In European Conference on Computer Vision (Springer), pp. 516–532.
26. Bernhardt, M., Castro, D.C., Tanno, R., Schwaighofer, A., Tezcan, K.C., Monteiro, M., Bannur, S., Lungren, M.P., Nori, A., Glocker, B., et al. (2022). Active label cleaning for improved dataset quality under resource constraints. Nat. Commun. *13*, 1161. https://doi.org/10.1038/s41467-022-28818-3.
27. Shen, Y., and Sanghavi, S. (2019). Learning with bad training data via iterative trimmed loss minimization. In International Conference on Machine Learning (PMLR), pp. 5739–5748.
28. Chen, Q., Jiang, G., Cao, F., Men, C., and Wang, W. (2024). A general elevating framework for label noise filters. Pattern Recogn. *147*, 110072. https://doi.org/10.1016/j.patcog.2023.110072.
29. Rokham, H., Pearlson, G., Abrol, A., Falakshahi, H., Plis, S., and Calhoun, V.D. (2020). Addressing inaccurate nosology in mental health: A multilabel data cleansing approach for detecting label noise from structural magnetic resonance imaging data in mood and psychosis disorders. Biol. Psychiatry. Cogn. Neurosci. Neuroimaging *5*, 819–832. https://doi.org/10.1016/j.bpsc.2020.05.008.
30. Cuthbert, B.N. (2014). The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. World Psychiatr. *13*, 28–35.
31. Parkes, L., Satterthwaite, T.D., and Bassett, D.S. (2020). Towards precise resting-state fMRI biomarkers in psychiatry: synthesizing developments in transdiagnostic research, dimensional models of psychopathology, and normative neurodevelopment. Curr. Opin. Neurobiol. *65*, 120–128. https://doi.org/10.1016/j.conb.2020.10.016.
32. Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P.Y., Cooper, J.L., Eaton, J., et al. (2018). The Lancet Commission on global mental health and sustainable development. Lancet *392*, 1553–1598. https://doi.org/10.1016/S0140-6736(18)31612-X.
33. Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M.R., and Rahman, A. (2007). No health without mental health. Lancet *370*, 859–877. https://doi.org/10.1016/S0140-6736(07)61238-0.
34. Lin, A., Reniers, R.L.E.P., and Wood, S.J. (2013). Clinical staging in severe mental disorder: evidence from neurocognition and neuroimaging. Br. J. Psychiatry. Suppl. *54*, s11–s17. https://doi.org/10.1192/bjp.bp.112.119156.
35. Sugranyes, G., de la Serna, E., Borras, R., Sanchez-Gistau, V., Pariente, J.C., Romero, S., Baeza, I., Díaz-Caneja, C.M., Rodriguez-Toscano, E., Moreno, C., et al. (2017). Clinical, cognitive, and neuroimaging evidence of a neurodevelopmental continuum in offspring of probands with schizophrenia and bipolar disorder. Schizophr. Bull. *43*, 1208–1219. https://doi.org/10.1093/schbul/sbx002.
36. Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research Domain Criteria (RDoC): Toward a new classification framework for research on mental disorders. Am. J. Psychiatry *167*, 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379.
37. Du, Y., Fu, Z., Sui, J., Gao, S., Xing, Y., Lin, D., Salman, M., Abrol, A., Rahaman, M.A., Chen, J., et al. (2020). NeuroMark: an automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders. Neuroimage. Clin. *28*, 102375. https://doi.org/10.1016/j.nicl.2020.102375.
38. Xia, S., Wang, G., Chen, Z., Duan, Y., and liu, Q. (2019). Complete random forest based class noise filtering learning for improving the generalizability of classifiers. IEEE Trans. Knowl. Data Eng. *31*, 2063–2078. https://doi.org/10.1109/TKDE.2018.2873791.
39. Zhou, Z.H. (2018). A brief introduction to weakly supervised learning. Natl. Sci. Rev. *5*, 44–53. https://doi.org/10.1093/nsr/nwx106.
40. Hinton, G.E., and Roweis, S. (2002). Stochastic neighbor embedding. Adv. Neural Inf. Process. Syst. *15*.
41. Andreasen, N.C., and Pierson, R. (2008). The role of the cerebellum in schizophrenia. Biol. Psychiatry *64*, 81–88. https://doi.org/10.1016/j.biopsych.2008.01.003.
42. Duan, M., Chen, X., He, H., Jiang, Y., Jiang, S., Xie, Q., Lai, Y., Luo, C., and Yao, D. (2015). Altered basal ganglia network integration in schizophrenia. Front. Hum. Neurosci. *9*, 561. https://doi.org/10.3389/fnhum.2015.00561.
43. Mueller, S., Wang, D., Pan, R., Holt, D.J., and Liu, H. (2015). Abnormalities in hemispheric specialization of caudate nucleus connectivity in schizophrenia. JAMA Psychiatr. *72*, 552–560. https://doi.org/10.1001/jamapsychiatry.2014.3176.
44. Ferri, J., Ford, J.M., Roach, B.J., Turner, J.A., van Erp, T.G., Voyvodic, J., Preda, A., Belger, A., Bustillo, J., O'Leary, D., et al. (2018). Resting-state thalamic dysconnectivity in schizophrenia and relationships with symptoms. Psychol. Med. *48*, 2492–2499. https://doi.org/10.1017/S003329171800003X.
45. Huang, H., Botao, Z., Jiang, Y., Tang, Y., Zhang, T., Tang, X., Xu, L., Wang, J., Li, J., Qian, Z., et al. (2020). Aberrant resting-state functional connectivity of salience network in first-episode schizophrenia. Brain Imaging Behav. *14*, 1350–1360. https://doi.org/10.1007/s11682-019-00040-8.
46. Rushworth, M.F., Ellison, A., and Walsh, V. (2001). Complementary localization and lateralization of orienting and motor attention. Nat. Neurosci. *4*, 656–661. https://doi.org/10.1038/88492.
47. Borgwardt, S.J., McGuire, P.K., Aston, J., Berger, G., Dazzan, P., Gschwandtner, U., Pflüger, M., D'Souza, M., Radue, E.W., and Riecher-Rössler, A. (2007). Structural brain abnormalities in individuals with an at-risk mental state who later develop psychosis. Br. J. Psychiatry *51*, S69–S75. https://doi.org/10.1192/bjp.191.51.s69.
48. Henseler, I., Falkai, P., and Gruber, O. (2010). Disturbed functional connectivity within brain networks subserving domain-specific subcomponents of working memory in schizophrenia: Relation to performance and clinical symptoms. J. Psychiatr. Res. *44*, 364–372. https://doi.org/10.1016/j.jpsychires.2009.09.003.
49. Du, Y., Fu, Z., Xing, Y., Lin, D., Pearlson, G., Kochunov, P., Hong, L.E., Qi, S., Salman, M., Abrol, A., and Calhoun, V.D. (2021). Evidence of shared and distinct functional and structural brain signatures in schizophrenia and autism spectrum disorder. Commun. Biol. *4*, 1073. https://doi.org/10.1038/s42003-021-02592-2.
50. Tamminga, C.A., Pearlson, G., Keshavan, M., Sweeney, J., Clementz, B., and Thaker, G. (2014). Bipolar and schizophrenia network for intermediate phenotypes: outcomes across the psychosis continuum. Schizophr. Bull. *40*, S131–S137. https://doi.org/10.1093/schbul/sbt179.
51. Keator, D.B., van Erp, T.G.M., Turner, J.A., Glover, G.H., Mueller, B.A., Liu, T.T., Voyvodic, J.T., Rasmussen, J., Calhoun, V.D., Lee, H.J., et al. (2016). The function biomedical informatics research network data repository. Neuroimage *124*, 1074–1079. https://doi.org/10.1016/j.neuroimage.2015.09.003.
52. Aine, C.J., Bockholt, H.J., Bustillo, J.R., Cañive, J.M., Caprihan, A., Gasparovic, C., Hanlon, F.M., Houck, J.M., Jung, R.E., Lauriello, J., et al. (2017). Multimodal neuroimaging in schizophrenia: description and dissemination. Neuroinformatics *15*, 343–364. https://doi.org/10.1007/s12021-017-9338-9.
53. Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B.S., Lewis, J.D., Li, Q., and

Milham, M. (2013). The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Front. Neuroinform. 7, 5.

54. Qi, S., Calhoun, V.D., van Erp, T.G.M., Bustillo, J., Damaraju, E., Turner, J.A., Du, Y., Yang, J., Chen, J., Yu, Q., et al. (2018). Multimodal fusion with reference: searching for joint neuromarkers of working memory deficits in schizophrenia. IEEE Trans. Med. Imaging *37*, 93–105. https://doi.org/10.1109/TMI.2017.2725306.
55. Du, Y., Fu, Z., and Calhoun, V.D. (2018). Classification and prediction of brain disorders using functional connectivity: promising but challenging. Front. Neurosci. *12*, 525. https://doi.org/10.3389/fnins.2018.00525.
56. Du, Y., and Fan, Y. (2013). Group information guided ICA for fMRI data analysis. Neuroimage *69*, 157–197. https://doi.org/10.1016/j. neuroimage.2012.11.008.
57. Wilson, D.R., and Martinez, T.R. (1997). Improved heterogeneous distance functions. J. Artif. Intell. Res. *6*, 1–34. https://doi.org/10.1613/jair.346.
58. Davies, D.L., and Bouldin, D.W. (1979). A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. *1*, 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.
59. Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
60. Caliński, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. Commun. Stat. Theor. Methods *3*, 1–27. https://doi.org/10.1080/03610927408827101.
61. Dunn†, J.C. (1974). Well-separated clusters and optimal fuzzy partitions. J. Cybern. *4*, 95–104. https://doi.org/10.1080/01969727408546059.
62. Vergara, V.M., Salman, M., Abrol, A., Espinoza, F.A., and Calhoun, V.D. (2020). Determining the number of states in dynamic functional connectivity using cluster validity indexes. J. Neurosci. Methods *337*, 108651. https://doi.org/10.1016/j.jneumeth.2020.108651.
63. Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods *8*, 665–670. https://doi.org/10.1038/nmeth.1635.
64. Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage *31*, 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015.
65. Ayachit, U. (2015). The Paraview Guide: A Parallel Visualization Application (Kitware, Inc.).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
BSNIP	Tamminga et al.[50]	https://doi.org/10.1093/schbul/sbt179
FBIRN	Keator et al.[51]	https://doi.org/10.1016/j.neuroimage.2015.09.003
MPRC	Du et al.[49]	https://doi.org/10.1038/s42003-021-02592-2
COBRE	Aine et al.[52]	http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html
ABIDEI	Craddock et al.[53]	https://fcon_1000.projects.nitrc.org/indi/abide/
Software and algorithms		
MATLAB R2018a	Mathworks	https://www.mathworks.com/
LAMP method	This paper	http://www.yuhuidu.com/index.php?a=cms&b=index&c=news&cid=183&id=8524
Complete random forest (CRF)-based label-noise filtering model	Xia et al.[38]	http://www.cquptshuyinxia.com/CRF-NFL.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yuhui Du (duyuhui@sxu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data from BSNIP, COBRE, and ABIDEI datasets are publicly available. Data from FBIRN and MPRC datasets are available from the corresponding author Yuhui Du (duyuhui@sxu.edu.cn) upon reasonable request due to privacy restrictions. DOIs/URLs are listed in the key resources table.
- All original codes that support the findings of this study have been deposited at Yuhui Du's personal website and are available from the corresponding author Yuhui Du (duyuhui@sxu.edu.cn) upon request. URLs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the corresponding author Yuhui Du (duyuhui@sxu.edu.cn) upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We analyzed fMRI data from 537 subjects with SZ and 708 age-matched HCs across four datasets, including BSNIP, FBIRN, MPRC, and COBRE, to execute our LAMP method. To validate the feasibility of our LAMP method on diverse populations, we used fMRI data from 398 subjects with ASD and 471 age-matched HCs from the ABIDEI dataset that were equally divided into four datasets, including SubData1, SubData2, SubData3, and SubData4. In this study, all data were approved by the local Institutional Review Board (IRB) committee, and the majority of participants were Caucasian. The corresponding demographic information, including age and gender, is detailed in Tables S1 and S30. These five datasets were employed from the following programs or institutes.

(1) Bipolar and Schizophrenia Network for Intermediate Phenotypes (BSNIP). BSNIP is from a publicly available database whose original source can be found at https://nda.nih.gov/, and the relevant reference of the data can be found at.[50]
(2) Function Biomedical Informatics Research Network Data Repository (FBIRN). The reference number of the IRB approval for FBIRN data is HS# 2009-7128.
(3) Maryland Psychiatric Research Center (MPRC). The reference number of the IRB approval for MPRC data is HP-00045716.
(4) Centers of Biomedical Research Excellence (COBRE). COBRE is from a publicly available dataset whose original source can be found at http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.
(5) Autism Brain Imaging Data Exchange (ABIDEI). ABIDEI is from a publicly available dataset whose original source can be found at https://fcon_1000.projects.nitrc.org/indi/abide/.

METHOD DETAILS

Participants and preprocessing

Resting-state fMRI data of HC and SZ subjects were selected from four multisite datasets, including BSNIP, FBIRN, COBRE, and MPRC. Data preprocessing and participant selection criteria followed our previous work.[49] After the quality control, the fMRI data of 537 subjects with SZ and 708 HCs from four different datasets were retained. Specifically, we used 182 SZ patients and 237 age-matched HCs from BSNIP, 137 SZ patients and 144 age-matched HCs from FBIRN, 150 SZ patients and 238 age-matched HCs from MPRC, and 68 SZ patients and 89 age-matched HCs from COBRE. Resting-state fMRI data of HC and ASD subjects were selected from ABIDEI dataset that were equally divided into four datasets, including SubData1, SubData2, SubData3, and SubData4. Specifically, we used 101 ASD patients and 116 age-matched HCs from SubData1, 103 ASD patients and 114 age-matched HCs from SubData2, 99 ASD patients and 119 age-matched HCs from SubData3, and 95 ASD patients and 122 age-matched HCs from SubData4. Detailed demographic information can be found in Tables S1 and S30.

We preprocessed the fMRI data using the statistical parametric mapping toolbox (SPM12). First, we removed the first six time points and performed rigid body motion correction to correct for subject head motion, followed by slice-timing correction to account for timing differences in slice acquisition. Next, we warped the fMRI data into the standard Montreal Neurological Institute (MNI) space using an echo-planar imaging template and resampled them to $3 \times 3 \times 3$ mm^3 isotropic voxels. To reduce noise, we smoothed the resampled fMRI images using a Gaussian kernel with a full width at half maximum (FWHM) of 6 mm. More specific information can be found in our previous work.[49]

Estimation of functional network connectivity (FNC) features from fMRI data

Brain FNC has been demonstrated the potential to reveal possible biomarkers for mental disorders.[54,55] In this work, we estimated the FNCs of fMRI data as initial features for each subject via a fully automated NeuroMark method[37] and toolbox (http://www.yuhuidu.com/ and http://trendscenter.org/software/gift) (see Figure 1A). The 53 intrinsic functional network templates were first constructed by identifying replicated and meaningful group-level independent components by independent component analysis (ICA) between two large healthy control datasets. Then, subject-level functional networks (FNs) and associated time courses (TCs) were automatically estimated for each individual subject using the group-information guided ICA (GIG-ICA)[56] by taking the templates as the guidance. After that, a static FNC matrix (53*53) was obtained for each subject by calculating the Pearson correlation coefficient between pairwise TCs of functional networks after postprocessing TCs.[37] Subsequently, we flattened the upper triangle of the FNC matrix into a vector, resulting in 1378 FNC measures for each subject. Moreover, we conducted a thorough regression analysis to carefully regress out the influences of age, gender, and site effects from the estimated FNC measures for each subject. Table S1 summarizes the demographic information for the participants.

Overall workflow of the present work

The overall workflow of the present work is displayed in Figure 1, which can be divided into two parts, including the construction of the LAMP method using fMRI data of HC-SZ (Figures 1A–1D and S1) and the evaluation of the constructed LAMP method (Figures 1E and 1F)). First, FNC features are estimated for each subject of the 1245 participants (mentioned in the previous sections), as shown in Figure 1A. In order to verify the result reproducibility, the four datasets are alternately divided into three source datasets and an independent dataset to perform the strict leave-one-dataset-out cross-validation procedure, as shown in Figure 1B. The three source datasets are used as training data to identify typical subjects from all subjects in each source dataset. The typical subjects are then used to construct a dimensional prediction model to predict the remaining independent dataset. Namely, all subjects in the remaining independent dataset are used as the test data to evaluate the performance of the prediction model. We identify typical subjects in the source datasets to construct the prediction model that provides dimensional scores and reassigned labels for subjects in the independent dataset. Specifically, as shown in Figure 1C, we employ the CRF label-noise filtering model to identify typical subjects whose labels align with the neuroimaging measures in each source dataset, which is the initial step to ensure the data validity of subsequent studies. More importantly, as shown in Figure 1D, we build a dimensional prediction model guided by the typical groups from the multiple source datasets to provide a dimensional score indicating the pathology and a new label for each independent subject. We evaluate the validity of the typical groups by analyzing the inter-group separability and intra-group compactness of the typical groups, as well as the significance and consistency of functional abnormalities within the typical SZ group across source datasets (see Figure 1E). As shown in Figure 1F, we evaluate the stability of the scores, inter-group separability and intra-group compactness of the relabeled groups, and the significant functional abnormalities within the relabeled SZ group in the independent dataset.

The proposed label-noise filtering-based dimensional prediction (LAMP) method

The proposed LAMP method includes two parts: (1) the identification of typical subjects in multiple source datasets (see Figure 1C) and (2) the construction of a dimensional prediction model for subjects in an independent dataset (see Figures 1D and S1). The method involves *n* source datasets and one independent dataset.

In the first part, we aim to identify the typical subjects of each source dataset based on a label-noise filtering model to ensure that biomarker discovery and disorder prediction are not confounded by the inconsistency between diagnostic labels and neuroimaging measures. Here, we employ the CRF-based label-noise filtering model[38] that assesses the label heterogeneity around each subject under various feature subspaces (i.e., feature subsets) to filter out subjects that it is surrounded by the heterogeneous subjects. On the one hand, the CRF-based model identifies subjects with high label heterogeneity in their surroundings as label noise based on the data itself rather than relying on classification results, thereby preventing different classifiers from detecting different label noise. On the other hand, the voting mechanism across various feature subspaces within the CRF-based model enables to process high-dimensional and noisy features effectively. However, a drawback of the CRF-based model is that subjects in the minority category tend to be mistakenly identified as label noise. Thus, before constructing the CRF-based model, we employ an under-sampling strategy to balance the category distribution and identify the typical subjects according to the non-noise ratio that describes the non-noise frequency to the sampling frequency of subjects. Specifically, the under-sampling CRF-based model consists of four steps to filter out label noise and identify the typical subjects for each source dataset.

Step 1: Under-sampling. To avoid biased results, we randomly select an equal number of subjects from each category to form a balanced dataset for identifying typical subjects.

Step 2: Construct a complete random forest. A complete random forest comprises N complete random decision trees (CRDTs) that are constructed following the rules below. The root node of each tree contains all of the sampled subjects. Non-leaf nodes in each tree have two child nodes, and the division rule of any node is based on randomly selecting both a feature and the split value of the feature. It is worth noting that the splitting process stops when the subjects in the node belong to the same category. The label of each node is determined by the majority of labels of the subjects in the node. The above process is repeated N times to form a complete random forest containing N CRDTs.

Step 3: Detect possible label noise in the constructed complete random forest. The label of a node that contains label noise is susceptible to change. Therefore, the stability of the node's label after the first change indicates the surrounding heterogeneity for each subject within the node and is used to detect label noise. Such stability is referred to as noise intensity (NI). The NI value for a subject corresponds to the number of times that the label of the leaf node containing the subject remains unchanged after the first change during traversing upward. If the NI value of a subject is greater than a given intensity threshold μ, the subject is provisionally considered a possible label noise by this CRDT. This complete random forest regards a subject as a label noise if more than 50% of the CRDTs in the forest detect it as a possible label noise.

Step 4: Detect typical subjects in multiple complete random forests. The above three steps are repeated t times to ensure that each subject is sampled at least once. If a subject is sampled T times and regarded as a possible label noise in T' complete random forest, the label-dependability of the subject can be calculated as $(T - T')/T$. Consequently, subjects whose label-dependability exceeds a predefined label-dependability threshold γ ($\gamma \in [0,1]$) are identified as typical subjects for subsequent analysis.

The CRF-based model involves three parameters: the number of repetitions denoted as t (which also represents the number of CRFs), the number of CRDTs denoted as N, and the intensity threshold denoted as μ. It is straightforward to comprehend that as the number of trees (N) and forests (t) increases, the detected noisy samples become more stable and reliable. In addition, the intensity threshold (μ) is closely related to the sample size. Essentially, with a larger μ, some label noise with weaker noise characteristics may be overlooked. Conversely, a smaller μ may result in some normal subjects being incorrectly labeled as label noise. Thus, considering the recommendations in the literature[38] and sample size, we set the t, N, and μ to 101, 200, and 2, respectively. Experimental results have shown that these parameters do not significantly affect the outcomes. The CRF-based model and under-sampling CRF-based model are described in detail in the literature[38] and Figure S1, respectively.

Furthermore, to evaluate the performance of the CRF-based model in filtering samples with label noise and identifying typical samples with matching labels and measures, we conduct an unbiased five-fold cross-validation classification experiment using three public datasets (details provided in Table S2). For each dataset, four folds of data are used as the training data and the remaining one fold is used as the test data. We introduce label noise by randomly flipping labels to different labels for samples in the training data. We conduct the CRF-based model to identify typical samples in the training data and then use these typical samples to construct a classifier. Test data is used to assess the performance of the classifier, which also indicates the model's capability to filter samples with label noise and identify typical samples. In this paper, we flip the labels of the training samples with noise rates of 5%, 10%, 15%, and 20%.

In the second part, we construct a dimensional prediction model using the typical subjects identified from the n source datasets to provide dimensional scores for the subjects in an independent dataset. In detail, we define a new similarity metric to assess the similarity between independent subjects and the two typical groups identified from each source dataset and construct the dimensional prediction model in light of the similarity. For each independent subject, we calculate $n+1$ scores, including n separate scores derived from each source dataset and one comprehensive score calculated by averaging the above n separate scores.

Let $\mathbb{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ represent the n source datasets. $\boldsymbol{G}^s_{HC}$ and $\boldsymbol{G}^s_{SZ}$ contain typical HCs and typical SZ patients selected from the source dataset $\boldsymbol{X}_s$, respectively, and $\boldsymbol{Z}$ denotes the independent dataset containing independent subjects. For each $\boldsymbol{G}^s_p$ ($p \in \{HC, SZ\}$), the corresponding center point ($\boldsymbol{C}^s_p$) is determined by the following two subject points:

(1) The subject point with the highest local density in group $\boldsymbol{G}^s_p$ is defined as

$$x^{\rho}_{G^s_p} = \max_{x \in G^s_p} \sum_{y \in Ne(x)} e^{-d(x,y)}, p \in \{HC, SZ\}, \quad \text{(Equation 1)}$$

where $d(\boldsymbol{x}, \boldsymbol{y})$ is Euclidean distance[57] between subject points $\boldsymbol{x}$ and $\boldsymbol{y}$, $Ne(\boldsymbol{x})$ contains the K nearest subjects closest to $\boldsymbol{x}$ in group $\boldsymbol{G}^s_p$. Here, we set K to five.

(2) The subject point with the farthest distance in group $\boldsymbol{G}^s_p$ to the different groups is defined as

$$x^{\delta}_{G^s_p} = \max_{x \in G^s_p} \sum_{y \in G^s_q} d(x,y), p, q \in \{HC, SZ\}, p \neq q. \quad \text{(Equation 2)}$$

The synthesized subject center point (C_p^s) is the weighted average of the above two subject points and is defined as

$$C_p^s = \omega_1 * x_{G_p^s}^{\rho} + \omega_2 * x_{G_p^s}^{\delta}, p \in \{HC, SZ\}. \quad \text{(Equation 3)}$$

For simplicity, we set balanced parameters ω_1 and ω_2 to 1/2 here.

Then, based on the center points C_{HC}^s and C_{SZ}^s derived from typical groups of the source dataset X_s, a separate score is computed for each independent subject by assessing its similarity to the two center points. Specifically, for any independent subject $z_k \in Z$, the similarity between z_k and C_p^s is

$$sim\left(z_k, C_p^s\right) = e^{-d\left(z_k, C_p^s\right)}, p \in \{HC, SZ\}. \quad \text{(Equation 4)}$$

The separate score indicating the relationship between independent subjects z_k and the two typical groups in X_s is designed in a straightforward formula by

$$DS(z_k, X_s) = sim\left(z_k, C_{SZ}^s\right) - sim\left(z_k, C_{HC}^s\right). \quad \text{(Equation 5)}$$

Consequently, we get n separate scores for the independent subject z_k, and the comprehensive score can be calculated by

$$Com_k = \frac{1}{n} \sum_{s=1}^{n} DS(z_k, X_s). \quad \text{(Equation 6)}$$

Those scores are normalized on a scale from -1 to 1, representing the continuum from no abnormality to significant abnormality. Specifically, the closer the score value is to -1, the greater the likelihood that the independent subject is normal. Conversely, the closer the score value is to 1, the higher the likelihood that the independent subject is schizophrenic.

To further verify the effectiveness of the dimensional prediction model, the independent subjects are relabeled by thresholding the comprehensive score via an adaptive parameter τ to obtain the relabeled HC group, relabeled SZ group, and the Boundary group in which subjects are middle and cannot be categorized into a specific group with enough confidence. Specifically, a truncation parameter ϵ of 0.2 is introduced to obtain the adaptive parameter τ that mitigates the scores bias of different groups in different datasets. To determine the value of τ, the comprehensive scores of all independent subjects are sorted in ascending order, and the number of scores less than 0 and greater than 0 are denoted as N^- and N^+, respectively. Then, the τ is defined by

$$\tau = \left| \frac{Com_{[N^- \times (1-\epsilon)]} + Com_{[N^- + N^+ \times \epsilon]}}{2} \right|, \quad \text{(Equation 7)}$$

where $[*]$ is the round function that replaces the number in square brackets with an approximate integer, and $|*|$ denotes the absolute value of a number. Consequently, the independent subjects are divided into three groups: the relabeled HC group comprises subjects with scores less than -τ, the relabeled SZ group comprises subjects with scores greater than τ, and the remaining subjects form the Boundary group.

Evaluation of the validity of the identified typical subjects in each source dataset

In this paper, we perform and evaluate the proposed LAMP method using four HC-SZ datasets, including three source datasets and an independent dataset. We conduct various experiments to demonstrate the improved inter-group separability and intra-group compactness of the typical groups identified from each source dataset (see Figure 1E). Besides, we explore significantly different FNC features between the typical HC group and typical SZ group and validate their consistency across multiple source datasets. It is worth noting that these experiments are performed between the original groups and between typical groups with various label-dependability thresholds in each source dataset.

First, to ensure the reliability of outcomes, we demonstrate that there are no significant differences in nuisance variables between typical groups (under different label-dependability thresholds) for each dataset. To do this, we evaluate the inter-group differences of the nuisance variables, including age and rotation/translations in the typical groups (under the different label-dependability threshold γ) by two-sample t-tests ($p < 0.01$ with Bonferroni correction). The Chi-square test ($p < 0.01$ with Bonferroni correction) is employed to evaluate the inter-group differences of gender in the typical groups (under the different label-dependability threshold γ). We also test these inter-group differences for original groups in the same way for comparison.

Second, we aim to confirm the enhanced inter-group separability and intra-group compactness of the identified typical groups compared with the original groups for each source dataset from four perspectives. Specifically, we first calculate the differences in P-values of all FNC features and the number of significant FNC features ($p < 0.01$ with Bonferroni correction) between original groups and between typical groups (under different label-dependability thresholds) via the two-sample t-test. Next, an unbiased and within-dataset 5-fold cross-validation classification framework is employed to investigate the separability between original groups and between typical groups (under different label-dependability thresholds) via classification accuracy. We divide the typical subjects into five folds, of which four folds are used as the training data to build classifiers, and the remaining one fold is used as the testing data to evaluate the classifiers. We also classify the original groups in the same framework. Here, we use five popular classifiers, including support vector machine with a linear kernel (SVM), decision tree (DT), 1-nearest neighborhood classifier (1NN), 3-nearest neighborhood classifier (3NN), and 5-nearest neighborhood classifier (5NN). We set the regularization parameter C in the SVM classifier to 100 and the parameter of the maximal number of decision splits per tree in DT to 1. In the following experiments, we set the label-dependability threshold to 0.8 for illustration. Additionally, we calculate five widely used evaluation metrics to measure inter-group separability and intra-group compactness in original groups and in typical groups. The five metrics include Davies Bouldin index (DBI),[58] silhouettes coefficient (SC),[59] Calinski-Harabasz index (CHI),[60] Dunn Validity index (DVI),[61] and the ratio of within-group similarity to between-group similarity (Sw/Sb) via Pearson correlation coefficients.[62] Moreover, we compare the inter-group overlap between typical groups and between original groups by visualizing the high-dimensional data in two dimensions using t-SNE projection technology.[40]

Third, we compute and analyze inter-group FNC differences within the typical groups in the source datasets to demonstrate the improved consistency and significance of patterns of the differences across datasets relative to the original groups. Here, we set the label-dependability threshold to 0.8 for illustration. In detail, we compute the T-values of FNC features ($p < 0.01$ with Bonferroni correction or without correction) between typical groups and between original groups in each source dataset based on the two-sample t-tests to demonstrate the significance of inter-group differences. Next, we compute the Pearson correlation coefficient for the T-values between typical groups and between original groups in paired source datasets to examine the consistency of inter-group differences across different datasets. Moreover, we exhibit the mean strength, P-values, and T-values of the top 10 FNC features that show the most significant inter-group differences in both the typical and original groups and the top 10 features that show significant inter-group differences only in the typical groups across source datasets.

Evaluation of the feasibility of the dimensional prediction model for the independent dataset

We conduct several experiments to demonstrate the feasibility of the proposed dimensional prediction model for independent subjects (see Figure 1F). To begin with, we calculate the Pearson correlation coefficients between the paired dimensional scores, including separate scores derived from three source datasets and comprehensive scores for the independent datasets to measure the model stability. A higher average correlation coefficient among these scores indicates that our model can obtain stable and consistent prediction results for the independent datasets, even using subjects from different datasets as guidance.

Then, we also assess the intra-group compactness, inter-group separability, and consistency of inter-group differences of the relabeled HC and SZ groups in the independent dataset via multiple evaluated experiments. First, using an unbiased 10-fold cross-validation classification framework, we divide the relabeled groups into ten folds, of which nine folds are used as the training data to build classifiers, including SVM, DT, 1NN, 3NN, and 5NN, and the remaining one fold is used as the testing data to evaluate the classifiers. We also classify the original groups in the same framework and compare the separability between original HC and SZ groups and between relabeled HC and relabeled SZ groups via classification accuracy, sensitivity, and specificity. Similarly, we calculate evaluation metrics DBI, SC, CHI, DVI, and Sw/Sb on the original groups and on the relabeled groups. Then, we project high-dimensional original groups and relabeled groups in a two-dimensional space based on t-SNE projection technology.

Furthermore, to demonstrate the enhanced inter-group differences within the relabeled groups, we computed the P-values and T-values of FNC features ($p < 0.01$ with Bonferroni correction or without correction) between relabeled groups and between original groups via the two-sample t-tests. Additionally, using the independent dataset, we show the mean strength, P-values, and T-values of the top 10 FNC features with the most significant inter-group differences shared between the original groups and relabeled groups and the top 10 unique FNC features with the most significant inter-group differences only in the relabeled groups.

Comparison of the typical subjects and the relabeled groups derived from the same dataset

In this work, each of the four datasets is successively used as the source dataset for typical subjects identification and prediction model construction and as the independent dataset for dimensional scores prediction and label reassignment via the leave-one-dataset-out division strategy. Since both typical subjects and relabeled subjects are demonstrated as valid subjects with matching labels and measures, we further confirm the validity of the LAMP method by discussing the consistency between typical subjects and relabeled subjects derived from the same dataset. In detail, for each dataset, we first construct a confusion matrix between the labels of typical groups and the relabeled groups derived from the same dataset to validate the consistency. Besides, we also explore stable impaired brain function in SZ patients as putative biomarkers. Specifically, to explore reliable, stable, and consistent impairments within SZ patients, we analyze the top 10 FNC features with significant inter-group differences that are shared in the original groups, typical groups, and relabeled groups across the four datasets based on two-sample t-tests. In the meantime, we explore the top 10 FNC features with significant inter-group differences observed in both the typical groups and relabeled groups across the four datasets, but these features are not significant within the original groups in any of the datasets. We outline the mean strength, between-group P-values, and between-group T-values of the 20 FNC features in these groups across the four datasets and visualize related functional networks in the brain via anatomical 3D surface meshes provided by Yushkevich[63] and three tools, including SPM (http://www.fil.ion.ucl.ac.uk/spm/), ITK-SNAP,[64] and ParaView.[65]

Reproducibility evaluation on SZ-HC data through the replication experiment

To further validate the stability and reproducibility of the findings obtained from our LAMP method, we conduct a replication experiment using an identical experimental workflow, as depicted in Figure 1. We assess the reliability and reproducibility of the proposed method by comparing the consistency between the results from two separate runs. Specifically, we also employ a strict leave-one-dataset-out cross-validation procedure, whereby each of the four datasets successively served as the independent dataset for the evaluation and the remaining three as the source datasets for constructing the dimensional prediction model. Typical HC and SZ subjects are identified from each source dataset by building a CRF-based label-noise filtering model. Then, using these typical subjects who present remarkable and consistent inter-group differences across source datasets, a dimensional prediction model is constructed to provide reliable dimensional scores indicating changes in brain function for independent subjects. We relabel the independent subjects to validate that the proposed model can also result in distinguishable groups. Based on the dimensional scores, the independent subjects are then categorized into three groups: the relabeled HC group, the relabeled SZ group, and the Boundary group in which subjects are in the middle and cannot be confidently categorized into a specific group. It is important to note that both the typical subjects in the source dataset and the relabeled subjects in the independent dataset have consistent labels with the fMRI measures. Using these subjects with consistent labels and fMRI measures, we also identify putative biomarkers in this run to reveal abnormal brain function in SZ patients compared to HCs. To evaluate the reproducibility of the outcomes obtained using our LAMP method, we assess the consistency of typical subjects across two runs for each source dataset, the consistency of relabeled subjects across two runs for the independent dataset, and the consistency of explored biomarkers across two runs.

Within and between datasets separability of the original groups and typical groups across SZ-HC datasets

To demonstrate the improved separability of typical groups within and between datasets compared with that of the original groups, we establish two different cross-validation classification frameworks. We first apply an unbiased 5-fold cross-validation classification framework to evaluate the separability of typical groups within each source dataset. Specifically, we divide the typical subjects in each source dataset into five folds, of which four folds are used as the training data to build classifiers, and the remaining fold is used as the testing data to evaluate the performance of the classifiers. This framework is also applied to classify the original groups in each source dataset. Furthermore, to evaluate the separability of typical groups between datasets among the three source datasets, we apply another unbiased leave-one-dataset-out cross-validation classification framework. In detail, typical subjects from one source dataset serve as the training data to build classifiers, and typical subjects from the remaining datasets are used as the testing data to evaluate the performance of the classifiers. This classification framework is similarly applied to the original groups from the source datasets. Here, we select typical subjects under the label-dependability threshold of 0.8 in each source dataset for illustration. Additionally, we employ five classifiers, including SVM, DT, 1NN, 3NN, and 5NN.

The generalizability validation of the proposed LAMP method on ASD-HC data

To validate the generalizability of our LAMP method on other disorders, we apply our LAMP method to fMRI data of subjects with ASD and age-matched HCs using the same experimental workflow (see Figure 1). More specifically, we utilize 398 subjects with ASD and 471 age-matched HCs, which were from the ABIDEI data. Here, we evenly divide the subjects into four datasets to thoroughly validate the generalizability of our work via a rigorous leave-one-dataset-out cross-validation procedure. For convenience, we call these four datasets as SubData1, SubData2, SubData3, and SubData4. Following the data processing procedure of our previous work,[49] we extracted 1378 functional network connectivity features for each subject, following a data processing procedure that carefully removed nuisance effects, including age, gender, motion, and site effects (see Table S30 for demographic information in detail). Similarly, we implement a rigorous leave-one-dataset-out cross-validation procedure, where each of the four datasets is sequentially used as the independent dataset for evaluation, while the remaining three datasets are utilized as the source datasets for developing the dimensional prediction model. For each source dataset, we identify the typical ASD and typical HC groups which are then used to construct a dimensional prediction model to provide reliable dimensional scores indicating changes in brain function for independent subjects. Similar to the evaluation of the LAMP method on SZ-HC data, we also assess inter-group separability, intra-group compactness, and inter-group differences of the typical ASD and typical HC groups for each source dataset. In addition, we assess the stability of the scores obtained from the dimensional prediction model, as well as the separability of the derived groups, i.e., relabeled ASD group and relabeled HC group. More importantly, using typical subjects and relabeled subjects who have matching labels and fMRI measures, we explore putative biomarkers that exhibit significant and consistent inter-group differences for ASD.

QUANTIFICATION AND STATISTICAL ANALYSIS

We assess inter-group differences in the nuisance variables, including age and rotation/translations, in the original groups, in the typical groups, and in the relabeled groups, using two-sample t-tests ($p < 0.01$ with Bonferroni correction). For evaluating inter-group differences in gender, we employ the Chi-square test ($p < 0.01$ with Bonferroni correction). Furthermore, to assess the significance of improved separability in the relabeled groups compared to the diagnostic groups, we use two-sample t-tests ($p < 0.01$) to evaluate the differences in classification performance between the two.