ARTICLE OPEN

Check for updates

Neuroimaging-based variability in subtyping biomarkers for psychiatric heterogeneity

Zhenfu Wen[1,2], Mira Z. Hammoud[1,2], Carole E. Siegel[1], Eugene M. Laska[1], Duna Abu-Amara[1], Amit Etkin[3,4], Mohammed R. Milad[1,2]✉ and Charles R. Marmar[1,5]✉

Neuroimaging-based subtyping is increasingly used to explain heterogeneity in psychiatric disorders. However, the clinical utility of these subtyping efforts remains unclear, and replication has been challenging. Here we examined how the choice of neuroimaging measures influences the derivation of neuro-subtypes and the consequences for clinical delineation. On a clinically heterogeneous dataset (total $n = 566$) that included controls ($n = 268$) and cases ($n = 298$) of psychiatric conditions, including individuals diagnosed with post-traumatic stress disorder (PTSD), traumatic brain injury (TBI), and comorbidity of both (PTSD&TBI), we identified neuro-subtypes among the cases using either structural, resting-state, or task-based measures. The neuro-subtypes for each modality had high internal validity but did not significantly differ in their clinical and cognitive profiles. We further show that the choice of neuroimaging measures for subtyping substantially impacts the identification of neuro-subtypes, leading to low concordance across subtyping solutions. Similar variability in neuro-subtyping was found in an independent dataset ($n = 1642$) comprised of major depression disorder (MDD, $n = 848$) and controls ($n = 794$). Our results suggest that the highly anticipated relationships between neuro-subtypes and clinical features may be difficult to discover.

Molecular Psychiatry (2025) 30:1966–1975; https://doi.org/10.1038/s41380-024-02807-y

INTRODUCTION

Psychiatric disorders labeled with a specific DSM diagnosis are often marked by wide heterogeneity in symptom profiles [1]. For example, post-traumatic stress disorder (PTSD), as defined in DSM-5, includes thousands of distinct patterns across reexperiencing, avoidance, mood and cognition, and hyper-arousal symptoms [2, 3]. At the neurobiological level, there has been an intense search for neuroimaging patterns that differentiate diagnosed cases from healthy controls. Much knowledge has been gained with these case-control studies towards understanding neurobiology in different psychiatric disorders, including but not limit to PTSD [4–6], major depression disorder [7, 8], and anxiety disorders [9, 10]. But limited insight has been gained towards translating this knowledge into clinical utility, either for advancing diagnostics or optimization of clinical care and outcomes. This gap can be attributed to many factors, one of the most prominent being the high heterogeneity of psychiatric disorders [11, 12]. Delineating homogeneous clusters from clinically heterogeneous samples and developing neurobiological markers for them could facilitate our understanding of psychopathology and advance precision treatment.

Because of this failure of translation of neurobiology to clinical outcomes, recent efforts have been refocused on delineating this connection in two ways. One is to search for clinical subtypes and relate these to underlying neurobiological mechanisms [13, 14]. In this approach, subtypes are often defined using the participants' clinical measures based on theory-driven methods (e.g., using psychological knowledges [13, 15]), or data-driven methods using clinical measures for clustering analysis [14]. The neurobiological differences between these defined subtypes are then examined. The second approach is to create biologically defined subtypes and then determine if these bio-subtypes are associated with clinically meaningful clusters of patients [11, 12]. While the second approach is often data-driven—using biological measures as clustering features to define subtypes, note that hybrid approaches which combining both theory-driven and data-driven approaches are also proposed in the literature [11, 16]. Recently, there has been an increase in efforts with emphasis on the second approach noted above; with studies combining neuroimaging data and machine learning tools to generate biologically-defined subtypes that would provide mechanistic explanations and differentiate patterns of clinical symptoms [11, 12, 17, 18]. Using structural and functional magnetic resonance imaging (fMRI) data, previous studies have identified neuro-subtypes of PTSD [19, 20], major depression disorder (MDD) [21, 22], schizophrenia [23, 24], and many other psychiatric disorders [12]. These identified neuro-subtypes were shown to exhibit distinct clinical/cognitive characteristics, facilitate case-control discrimination, or respond differently to treatments, suggesting potential clinical utilities of the neuroimaging-based

[1]Department of Psychiatry, Grossman School of Medicine, New York University, New York, NY, USA. [2]Faillace Department of Psychiatry and Behavioral Sciences, McGovern Medical School, University of Texas Science Center at Houston, Houston, TX, USA. [3]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. [4]Alto Neuroscience, Mountain View, CA, USA. [5]Neuroscience Institute, New York University, New York, NY, USA. ✉email: Mohammed.r.milad@uth.tmc.edu; Charles.marmar@nyulangone.org

Received: 28 February 2024 Revised: 15 October 2024 Accepted: 18 October 2024
Published online: 7 November 2024

subtyping. Despite this progress, however, promising reports of clinically meaningful neuro-subtypes of MDD [21], PTSD [19], and trauma-related resilience [25] were not replicated, or only partially replicated, across studies. For example, in multiple attempts of conceptional nonexact replications, Dinga et al. [26] did not replicate MDD subtypes reported in Drysdale et al. [21], Esterman et al. [27] did not replicate PTSD subtypes reported in Etkin et al. [19], and Ben-Zion et al. [28] did not replicate trauma-related resilience subtypes reported in Stevens et al. [25].

Identifying and reproducing clinically meaningful subtypes across studies can be challenging. On the one hand, the identified subtypes may not be internally validated and thus hard for external generalization. Internal validation is critical for clustering analysis since most clustering algorithms will generate cluster solutions even when there are no underlying cluster structures [26, 28, 29]. Some replication studies have demonstrated that statistically significant subtypes may not be identified using similar methodology as the original studies [26, 28]. Even if significant subtypes are identified, further validations about the clinical utility are required. Many studies identified subtypes with distinct biological measures (which is expected since the subtypes were defined based on these measures) but similar phenotypic profiles (e.g., see [12] for review). And there is a trend that studies demonstrating clinical utility tended to use less stringent reproducibility validation strategy [12], suggesting the complexity of the identification of clinically meaningful neuro-subtypes.

Another factor that can contribute to the difficulties in replications is the variability of biological features used for the clustering analysis [18, 30]. Different imaging modalities are commonly collected in neuroimaging studies. And different feature types from the same modality can be ascertained as clustering features. For example, cortical thickness or brain volumes from structural images, different white matter metrics from diffusion tensor imaging, functional connectivity or Regional Homogeneity (ReHo) [31] from resting-state fMRI data, activation patterns of different contrasts from task fMRI data. In the subtyping literature, a variety of features were used across studies, usually without clear justifications [12]. These differences of clustering features make it hard to compare results across studies. How can we tell whether subtypes reported in two different studies are capturing similar axis of heterogeneity of a psychiatric disorder? To answer this question, it is important to know to what degree does the choice of modalities and feature types impact the identification of subtypes. One may argue that it is a truism that different modalities and different measures from the same modality have different clustering solutions. However, there is clear evidence demonstrating links between different modalities, e.g., structure-function coupling [32], and similarities between measures derived from the same modality [33]. Therefore, it is nontrivial to examine the concordance between subtypes identified using different features. From the view of reproducibility, a low concordance between subtyping solutions will inform us to match clustering features as much as possible in replication studies. However, to our knowledge, there is no study that has systematically examined the impact of imaging features on the identification of subtypes within the same sample.

In this study, we aimed to provide answers to the following questions: 1) could internally valid subtypes be defined for different neuroimaging modalities? 2) Do the modality-specific subtypes have different clinical and cognitive profiles? 3) Is there concordance in the identified subtypes? And 4) do alternative clustering methods lead to similar observations? To answer these questions, we conducted multiple analyses on a clinically heterogeneous neuroimaging dataset ($n = 566$) to examine how the choice of neuroimaging measures influences the derivation of subtypes and investigated the clinical characteristics of the identified subtypes. We performed similar analyses on an independent imaging dataset ($n = 1642$) of participants with Major Depressive Disorder (MDD) to examine whether the answers to the questions listed above on this dataset are similar with those we observed on the main dataset.

MATERIALS AND METHODS

Participants

We analyzed a clinically heterogeneous neuroimaging dataset that includes a total of 566 participants from the NYU Cohen Veterans Center cohort [19, 20]. The dataset included individuals with no current psychiatric diagnosis (healthy controls, $n = 268$) and cases ($n = 298$) diagnosed with the trauma-related conditions of PTSD ($n = 79$), traumatic brain injury (TBI, $n = 168$), or comorbidity of both (PTSD&TBI, $n = 51$). The participants were recruited at New York University ($n = 356$) and Stanford University ($n = 210$). All participants were combat veterans. All participants provided informed consent before their participation and procedures were approved by both institutions' Institutional Review Boards. Subsets of the dataset were used in our previous studies with different research objectives. Specifically, Etkin et al. [19] used the data as a validation set for the identified memory-deficit PTSD subgroup, and Maron-Katz et al. [20] used the data to define PTSD subtypes using abnormal resting-state functional connectivity. The objective of this study is different from the previous studies. Here, we examined the validities of different modality-specific subtypes, and impact of analysis choices on the identified subtypes.

Neuroimaging data acquisition and preprocessing

Participants underwent multiple scanning runs including a resting-state run, a task-fMRI run, and a structural run. During the resting-state run (which lasted 8 min), the participants were instructed to remain awake and look at a fixation on the screen. During the task-fMRI run, the participants underwent a well-established emotional conflict paradigm [34, 35]. During the structural run, a high-resolution T1-weighted structural scan was acquired. Participants were scanned either using a 3.0 Tesla Siemens Magnetom Skyra scanner at NYU or a GE 750 scanner at Stanford University using the same scanning parameters (see Supplementary Materials). The structural images were preprocessed using the Computational Anatomy Toolbox (CAT12) [36], and the functional images were preprocessed using fMRIPrep 20.0.2 [37] (see Supplementary Materials).

Feature extraction for clustering

Structural measures. Regional volumes were extracted using the neuromorphometrics atlas in CAT12. This atlas comprises 134 regions of interest (ROIs) from grey matter, white matter, and cerebrospinal fluid. The total intracranial volume (TIV) was regressed out from the regional values. These extracted values were concatenated to a 134-dimensional vector for clustering analysis.

Task-fMRI measures. Participants underwent a validated emotional conflict paradigm during fMRI scanning [34, 35]. The contrast map of conflict detection was used for the clustering analysis (see Supplementary Materials). The contrast was defined as the incongruent trials ('I') versus congruent trials ('C'). To reduce the feature number, we used regional activations instead of voxel activations as features. A 442-region whole-brain parcellation consisting of 400 cortical regions [38], 32 subcortical regions [39], and 10 cerebellum regions [40] was used to extract regional activation measures. We used this combination of atlases because they provide good spatial resolution in reflecting the functional boundaries across brain regions and are widely used in the literature. Therefore, a 442-dimensional vector from each subject was used for clustering.

Resting-state measures. For each participant, the mean time series were extracted based on the above mentioned 442-region whole-brain parcellation. Pearson correlation was calculated between the time series of every two regions, which resulted in a 442×442 functional connectivity (FC) matrix for each participant. The matrixes were transformed using Fisher's r-to-z transformation and then averaged across rows to obtain regional connectivity values. This resulted in a 442-dimensional vector representing a whole-brain resting-state FC pattern for each participant, which was used for clustering.

Covariates correction. Since the neuroimage data were collected from two different sites, we further used the ComBat harmonization method [41] to

correct the effect of sites on each data modality. The ComBat harmonization method is widely used for correcting site effects in neuroimaging studies. It was initially introduced for diffusion tensor imaging measures [41] and subsequently been applied to structural and functional MRI data [42, 43]. ComBat harmonization was separately conducted for structural, task-fMRI, and resting-state measures across the entire dataset (i.e., including all controls and cases). The biological variates that were protected for during the removal of site effects including diagnosis, age, and sex. An empirical Bayes procedure with parametric priors was performed for the harmonization model estimation. We also regressed out the age and sex effects from the features before conducting clustering analysis.

Subtyping with K-means clustering

K-means. The K-means clustering algorithm from scikit-learn toolbox [44] was used to cluster participants into subtypes based on their structural measures, task-fMRI measures, or resting-state measures. Clustering was separately run on features from each modality, with the cluster number k varying from 2 to 9. The final cluster number for the analysis was determined using the Calinskin-Harabasz score and silhouette score [45].

Significance of the subtypes. The K-means algorithm will generate a clustering solution even if there is no underlying cluster structure. Thus it is important to test if the identified subtypes can be observed in unstructured data. We used the SigClust approach proposed by Liu et al. [29] to test the significance of the subtypes. This approach uses a Monte Carlo procedure to test if the observed data can be modeled as coming from a single multivariate Gaussian distribution. This method assumes under the alternative that the data are distributed as a mixture on multivariate normal distributions, each component of the mixture corresponding to a cluster. The null hypothesis is that the data is not a mixture- it is a single multivariate normal distribution. Failing to reject the null implies there are no underlying clusters.

Stability of subtypes. A stable clustering solution should be robust to small perturbations of the dataset, such as removing a subset of subjects from the clustering procedure. We evaluated the stability of the subtypes by resampling the data, reidentifying the clusters, and examining the effect of the data perturbation on the cluster memberships. Specifically, we randomly selected 80% of participants from the whole dataset, and rerun the K-means clustering with the same settings to cluster participants into 2 subtypes. We repeated the procedure 100 times to produced different subtyping solutions. We then calculated the adjusted rand index (ARI) and adjusted mutual information (AMI) between every two subtyping solutions within the participants that occurred in both resampled sets. The ARI and AMI are standardized measures usually ranging from 0 to 1. The values of both measures that are close to 1 represents a high concordance between the two clustering solutions, i.e., two individuals in the same cluster in one clustering solution have a high probability of being together in the other clustering solution. The values close to or smaller than 0 represent that the two clustering solutions do not have this property. Note that the ARI and AMI values can be related to the degree of data perturbation, where larger data perturbation usually results in lower ARI/AMI values.

Separability of clusters. We used a recently proposed selective inference approach [46] to test the separability of the identified subtypes. This approach tests the difference in means between clusters identified via K-means. A significant difference in means between subtypes suggests that the subtypes are distinct from each other in their neuroimaging patterns. It is biased to test the difference in means of clusters using classical statistical methods because the clusters are not random samples from predefined populations. Rather, they are produced by the K-means algorithm. The Chen and Witten method [46] corrects for this issue.

Feature differentiation. We conducted a post-clustering difference test to examine the subtype differences feature by feature, i.e., which features were significantly separate the identified subtypes from each other for each modality. A classical statistical test, e.g., two-sample t-test, is also biased in this situation, because the double use of the data (features were used for clustering and feature-level statistical test) leads to the failure of controlling the Type I error rate. We used a recently proposed post-clustering difference testing procedure that accounts for the clustering process, and better controls the Type I error rate than classical statistical methods [47].

Differences in clinical and cognitive measures between subtypes

We examined 31 different psychometric measures, including 20 clinical-related measures, and 11 neurocognitive measures estimated using participants' behavioral metrics across multiple tasks [48]. We used the two-sample t-test to compare the clinical and cognitive measures between subtypes. Different from the neuroimaging features, the clinical and cognitive measures were not used for defining the subtypes, the two-sample test was non-biased here. Multiple comparisons were corrected using the false discovery rate (FDR) correction. The confidence intervals of effect sizes (Cohen's d) were estimated using bootstrap resampling (1000 times).

We further examined the clinical/cognitive profiles using continuous subtyping assignments, a strategy similar to previous study [49]. The rationale behind this analysis is that continuous assignments may better capture individual variations than discrete assignments of individuals to one of the subtypes. For example, for individuals whose features lie near the boundary of the two different subtypes, a small change of their feature patterns may alter their discrete subtype assignments. In contrast, the continuous assignment scores which measure the similarities between these feature patterns and the subtype centroids will remain relatively stable. The procedure of this analysis is as follows. First, we obtained the mean pattern of each subtype (i.e., centroid of each clustering group) using K-means clustering. Second, we calculated Pearson's correlation between each participant's feature pattern (e.g., regional volumes for structural-based clustering) and the centroid of each subtype. This provided a subtype score that quantified the similarity between the participant's feature pattern and the corresponding subtype pattern/centroid. The subtype score is a continuous measure ranges from -1 to 1, where -1 indicates the participant's feature pattern is totally different from the subtype centroid, and 1 indicates a perfect match with the subtype centroid. Since we identified two subtypes for each modality, we obtained 2 subtype scores (one for subtype 1, one for subtype 2) for each participant. Third, we calculated the correlation between each subtype score and clinical/cognitive measure across participants. A significant correlation indicates that similarity to the subtype is associated with the corresponding clinical/cognitive measure. We separately conducted the above procedure for each modality.

In addition to analyzing clinical and cognitive profiles between subtypes, we also examined whether subtyping could enhance performance in distinguishing cases from healthy controls using neuroimaging measures (See Supplemental Materials for details).

Consistency of subtypes defined using different feature types

For participants that were assigned to the same cluster/subtype based on one modality (e.g., structural volume), we examined the percentage of them that were still assigned to the same cluster based on another modality (e.g., task-fMRI activation, or resting-state FC). We further calculated the ARI and AMI between clustering solutions obtained using features from two different modalities. A high ARI/AMI value (close to 1) indicates that participants were similarly clustered across modalities.

Similar analyses were conducted between clustering solutions based on different features extracted from the same data modality. For the structural data, instead of using the regional volumes extracted via CAT12, we used another feature type that was commonly used in the literature—cortical thicknesses extracted via Freesurfer 5.0 to define two subtypes. For the task-fMRI activation data, we used another activation contrast that was used to examine neural mechanisms of conflict resolution for the clustering. The contrast was defined as the incongruent trials preceded by incongruent trials ('il') versus incongruent trials preceded by congruent trials ('cl'). For resting-state FC, we used another nuisance regression strategy before connectivity estimation. In addition to the nuisance regressors used in the main analysis, we included the mean whole-brain signal as a nuisance regressor in the analysis. The inclusion of this global signal regression (GSR) step is debated in resting-state fMRI literature [50, 51], while many studies suggested that GSR may provide advantages over preprocessing without GSR (e.g., [52, 53]).

Similar analysis on another dataset

We conducted additional analyses on another large neuroimaging dataset from the REST-meta-MDD Project [54]. The objective of the analysis was not to directly compare subtypes identified from different datasets or disorders, but to examine if the main observations from the main dataset still hold on another dataset. These observations include: 1) limited

difference in clinical measures between subtypes; 2) low concordances between clustering solutions using different features. We used the REST-meta-MDD dataset because 1) The heterogeneity of MDD is well-documented and there are many subtyping studies for MDD; 2) The REST-meta-MDD dataset contains a large number of participants; 3) The REST-meta-MDD dataset provides data preprocessed with a standardized pipeline, significantly reducing the computational load.

We focused on the participants used in a previous study [55], which including 848 MDD patients and 794 individuals without MDD (health controls). Since no task-based fMRI were available, we only focused on examining the impact of different features estimated using the resting-state fMRI data. Future studies should examine the utility of different imaging modalities. Three different kinds of features were extracted for the clustering analysis: functional connectivity (Conn), ReHo [31], and fractional amplitude of low-frequency fluctuations (fALFF) [56]. The Harvard-Oxford atlas was used to extract regional feature values for the K-means clustering. Note that we did not use the 442-region atlas as in the main dataset because the REST-meta-MDD dataset only provided estimated functional connectivity matrixes rather than raw data, so we could not use the 442-region atlas. Only data from the MDD patients were used in the clustering analysis. We identified two subtypes based on Conn, ReHo, or fALFF features, respectively. The cluster number was set to two because a previous subtyping study [22] on the REST-meta-MDD dataset suggested that two was the optimal cluster number. For clinical measures, we focused on the Hamilton Depression Rating Scale (HAMD), a widely used depression assessment scale. The HAMD was the primary clinical measure in the REST-meta-MDD dataset, with data available from the majority of MDD patients. We also examined the Hamilton Anxiety Rating Scale (HAMA) and illness duration for a subset of patients whose data were available.

RESULTS

Can internally valid subtypes be defined for different neuroimaging modalities?

We first examined if internally valid bio-subtypes could be derived from each of those neuroimaging modalities: 1) structural data, 2) task-fMRI data, and 3) resting-state fMRI data. We constructed the subtypes based on the cases population ($n = 298$) as most of previous unsupervised subtyping studies did [12]. We set the cluster number to 2 as suggested by the Calinskin-Harabasz score and silhouette score (Figure S1). The K-means algorithm identified two subtypes for structural features (labeled S1 and S2), task-fMRI features (labeled T1 and T2), and resting-state features (labeled R1 and R2), respectively. These modality-specific subtypes were comprised of a similar proportion of the three diagnostic groups (PTSD, TBI, or PTSD&TBI; Fig. 1A-C): S1 (28%, 52%, 20%),

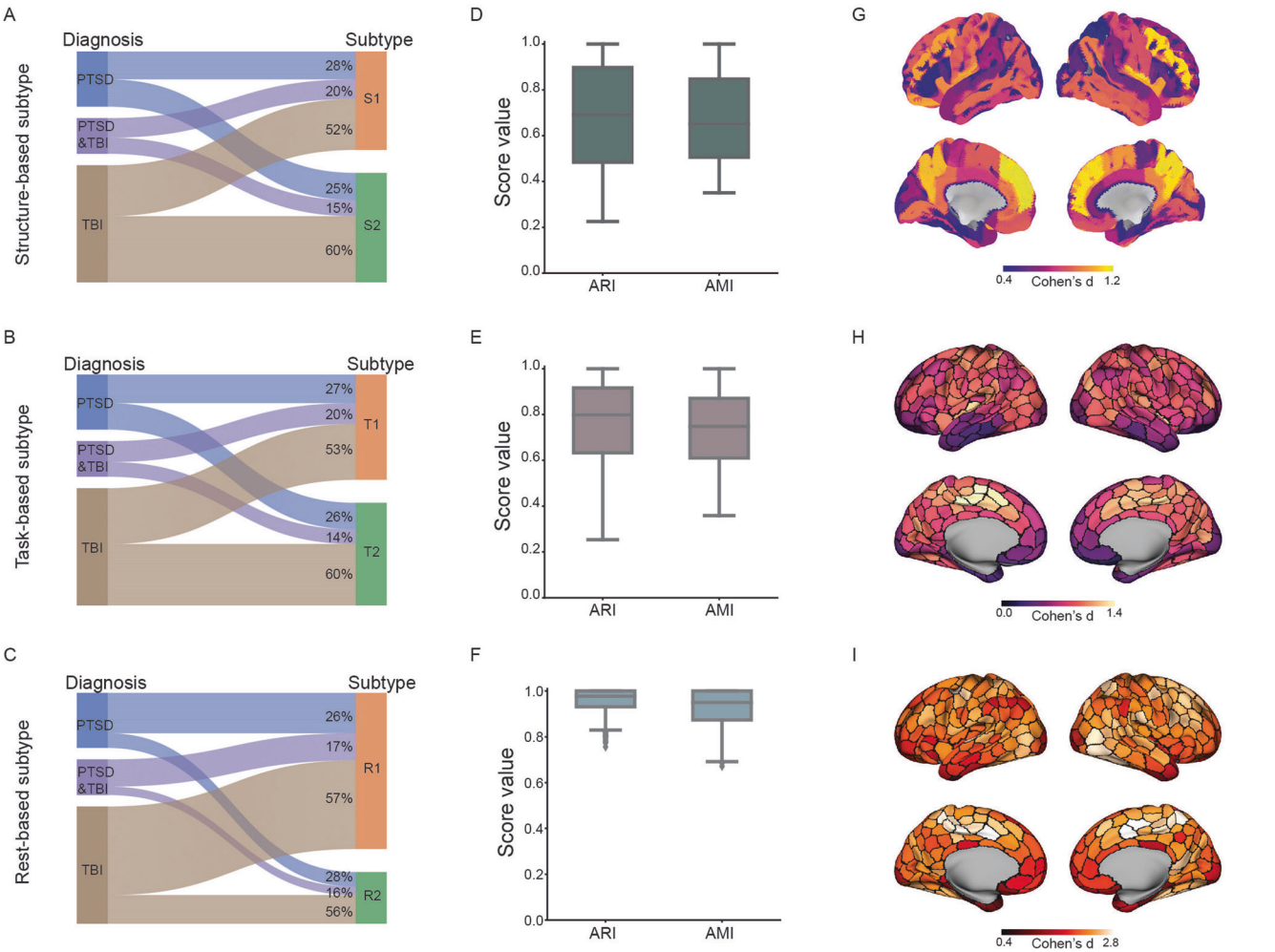

Fig. 1 Subtypes identified use a single modality of neuroimaging data. A–C Diagnoses distribution across the modality-specific subtypes. For each modality, the Sankey diagram depicts subtype assignments for participants from the three diagnose groups (PTSD, TBI, and PTSD&TBI). **D–F** Stability measures of the modality-specific subtypes. For each modality, the adjusted rand index (ARI) and adjusted mutual information (AMI) values were estimated between clustering solutions on resampled data (80% participants, without replacement, 100 times). Bounds of the box represent the 1st (25%) and 3rd (75%) quartiles, the central line represents the median, the whiskers represent the values within 1.5 times of the interquartile range, the flier points represent outliers falling beyond the whiskers. **G-I**. Differences in the modality-specific patterns between the two subtypes. For each regional measure, the statistical difference between the two subtypes was tested using the post-clustering difference testing procedure proposed by Hivert et al. [47]. The effect size (Cohen's d) of each region was reported.

S2 (25%, 60%, 15%), T1 (27%, 53%, 20%), T2 (26%, 60%, 14%), R1 (26%, 57%, 17%), R2 (28%, 56%, 16%).

Next, we examine the internal validity of the identified modality-specific subtypes from four different aspects. First, we assessed the 'significance of clustering' using the SigClust approach [29]. This analysis showed a preference for the existence of subtypes for all 3 modalities ($p_{FDR}< 0.001$). Second, we tested the stability (or reproducibility) of the subtypes using a resampling and re-clustering approach. All three modalities resulted in high ARI and AMI across the resampling procedure (Fig. 1D–F), suggesting that the constituents within each of the subtypes are stable under perturbations to the data. Note that stability does not imply separability, it is possible to get very stable clustering solutions from continuous data that are not separatable. Third, we tested the separability of clusters using a selective inference approach [46]. This test showed that the multivariate mean vectors of the two clusters were significantly different for all 3 modalities (structure-based: $p_{FDR} = 0.0015$; task-based: $p_{FDR} = 4.7$ E-25; rest-based: $p_{FDR} = 7.4$E-70). Fourth, we examined the feature differentiation between subtypes using a post-clustering difference testing [57]. This analysis demonstrated that the modality-specific subtypes significantly differed across most of the regions examined ($p_{FDR}< 0.05$; 90% regions for S1 vs. S2; 99% regions for T1 vs. T2; 100% regions for R1 vs. R2; Fig. 1G–I). Overall, these analyses suggest that the identified neuro-subtypes are internally valid and likely reflect the heterogeneity of the neuroimaging data. Note that because of the limitations of the testing methods, e.g., required assumptions may not be fully met, these analyses only provided evidence but do not guarantee the separability of the identified subtypes. Internal validations of clustering results are challenging in settings without ground truth, the external validations with measures not used in the clustering analysis are needed.

Do the modality-specific subtypes have different clinical and cognitive profiles?

We compared 31 different clinical and/or cognitive features between modality-specific subtypes. For the structure-based subtypes (Figure S2A), we found significant differences between S1 and S2 in their Beck Depression Inventory scores (BDI, $p_{uncorrected} = 0.024$) and Emotion Identification scores ($p_{uncorrected} = 0.039$). For the task-based subtypes (Figure S2B), we found significant differences between T1 and T2 in physic abuse scores on the Early Trauma Inventory (ETI_Phy_Abuse, $p_{uncorrected} = 0.026$). For the rest-based subtypes (Figure S2C), we found significant differences between R1 and R2 in their total scores of Early Trauma Inventory-Self Report (ETISR, $p_{uncorrected} = 0.046$), general trauma scores on the Early Trauma Inventory (ETI_Gen_Trauma, $p_{uncorrected} = 0.005$), BDI ($p_{uncorrected}=0.037$), and Sustained Attention scores ($p_{uncorrected}=0.045$). However, none of the above-mentioned measures survived correction for multiple comparisons ($p_{FDR}> 0.10$). Comparing the measures between subtypes only within individuals with PTSD or TBI resulted in similar findings ($p_{FDR}> 0.10$).

We further explored if continuous rather than discrete subtype assignments better capture the associations between neuroimaging data and clinical/cognitive profiles (Figure S2D). In this analysis (Figure S2E), we found significant correlations between S1-score and BDI ($r = -0.14$, $p_{uncorrected} = 0.043$), T1-score and Emotion Identification score ($r = 0.13$, $p_{uncorrected} = 0.033$), R2-score and Flexibility score ($r = 0.16$, $p_{uncorrected} = 0.010$). But as before, none of the above-mentioned measures survived correction for multiple comparisons. ($p_{FDR}>0.10$). We conducted a similar analysis across the whole sample (all cases). Specifically, we calculated the centroid by averaging feature patterns of the whole sample, and then calculated Pearson's correlation between the centroid and each participant's feature pattern. This similarity score was used to assess its correlation with clinical/cognitive measure across participants. The analysis did not reveal significant correlations for any modality (all p-values > 0.05 without multiple comparison corrections). In addition, the classification analyses suggested that subtyping did not improve the ability to distinguish cases from controls (Figure S3).

Is there concordance in the identified subtypes?

First, we examined whether the participants clustered into the same subtype in one modality (e.g., structural measure) were also clustered together in another modality (e.g., task-fMRI). The compositions of the subtypes were quite different across modalities (Fig. 2A). For example, 44% and 56% of individuals in task-based subtype 1 (T1) were clustered as structure-based subtype 1 (S1) and structure-based subtype 2 (S2), respectively. And Individuals in rest-based subtype (R1) were almost evenly composed of individuals from T1 (50%) and T2 (50%). To quantitatively measure the concordance of different modality-specific subtypes, we calculated the ARI and AMI between subtype assignments across every two modalities. We observed that both ARI and AMI values were near zero in all scenarios (Fig. 2B), suggesting low concordance of the identified subtypes across modalities. The results remained consistent when using cortical thickness instead of regional volume to identify structure-based subtypes (Figure S4).

Second, we further investigated how the choice of features within a modality could impact the subtyping results. For the structural data, instead of using the regional volumes, we used cortical thicknesses to define two subtypes (labeled St1 and St2). Interestingly, the components of St1 and St2 did not match the subtypes defined using regional volumes (S1 and S2), with St1 (St2) composed of 47% (38%) individuals from S1 and 53% (62%) individuals from S2 (Fig. 2C). The ARI and AMI measures between the two different subtyping solutions were less than 0.1, indicating low concordance. Similarly, for subtyping using task activations, a different task contrast from the same task paradigm also led to different subtypes (labeled Tr1 and Tr2), as demonstrated by the low to moderate ARI and AMI measures (< 0.25, Fig. 2D). For subtyping using resting-state functional connectivity, we again observed inconsistent subtypes across the two processing options (with or without GSR), with the ARI and AMI measures below 0.1 (Fig. 2E). These results suggest that even using data from the same neuroimaging modality, selecting different feature types as input for the cluster algorithm could lead to very different subtypes.

Do alternative clustering methods lead to similar observations of limited concordance across subtypes?

We used K-means as the main clustering algorithm. Although K-means is widely used in neuroimaging-based subtyping, a downside of this approach is that it does not incorporate the diagnosis information into the model. To address this limitation and potentially identify diagnosis-related subtypes, we conducted additional analyses using a recently proposed semi-supervised clustering method—Heterogeneity through Discriminative Analysis (HYDRA) [58, 59]. Instead of directly clustering cases based on their neuroimaging data, HYDRA clusters cases by maximizing the separation between healthy controls and the case subtypes. We conducted HYDRA-based clustering on each of the three modalities (see Supplementary Material). As shown in Fig. 3A, HYDRA achieved moderate to high ARI/AMI across the resampled data, suggesting stable subtyping solutions. In addition, HYDRA-based subtypes showed moderate to high concordance to those subtypes identified using K-means (ARI/AMI > 0.45, Fig. 3B), suggesting good consistency across clustering methods within each modality. However, similar to the solutions of K-means, the between-modality consistencies of the HYDRA-based subtypes were low between every two modalities (ARI/AMI < 0.1, Fig. 3C). We did not observe significant differences in clinical/cognitive measures between subtypes identified using each of the modality data after multiple comparison corrections (all $p_{FDR} > 0.10$).

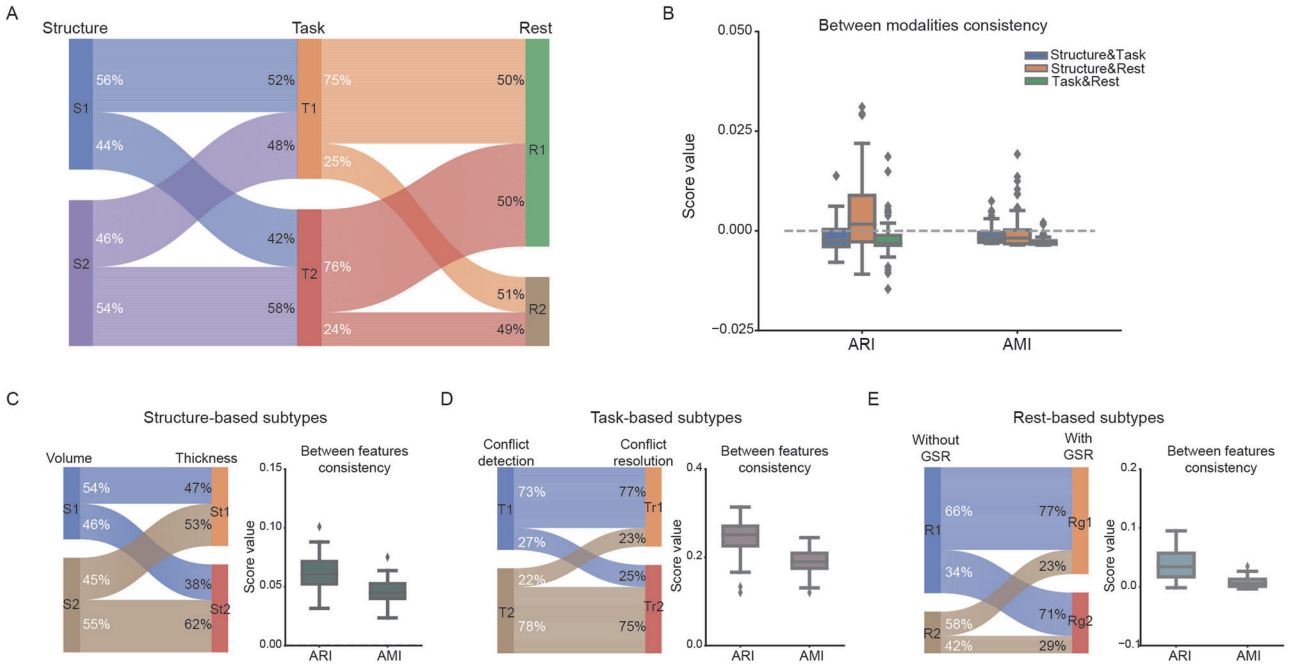

Fig. 2 Low consistency of subtyping solutions between and within data modalities. A Subtype solutions across modalities. The Sankey diagram depicts different subtype assignments across clustering solutions for the three data modalities. **B** Between modalities consistency of the identified subtypes. The adjusted rand index (ARI) and adjusted mutual information (AMI) values were estimated between clustering solutions derived using different data modalities. **C**–**E** Consistency of the subtypes using two different feature types from the (**C**) structural, (**D**) task-based, or (**E**) resting-state data. For each data modality, the ARI and AMI values were estimated between clustering solutions derived using two different clustering features. Bounds of the box represent the 1st (25%) and 3rd (75%) quartiles, the central line represents the median, the whiskers represent the values within 1.5 times of the interquartile range, the flier points represent outliers falling beyond the whiskers.

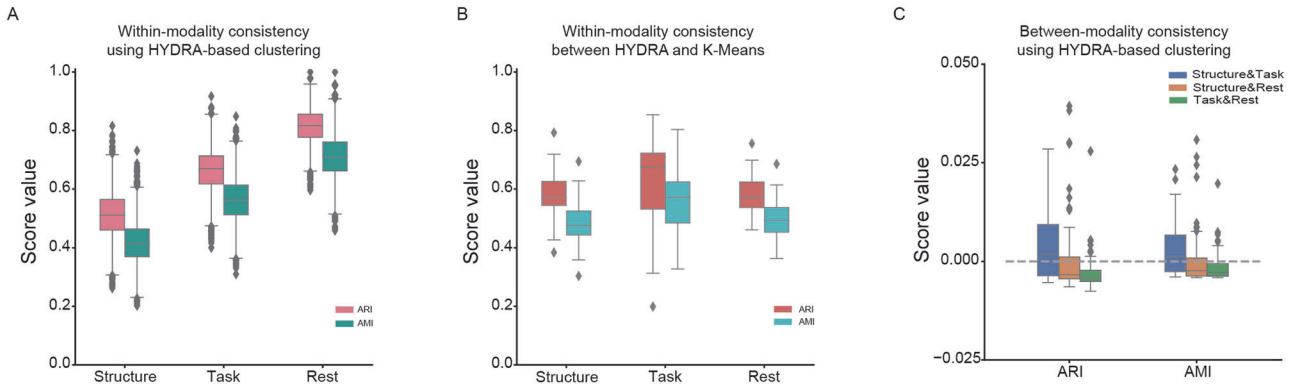

Fig. 3 Subtyping using alternative methods. A Consistency of subtyping solutions for each modality of data by using Heterogeneity through Discriminative Analysis (HYDRA). For each modality, the adjusted rand index (ARI) and adjusted mutual information (AMI) values were estimated between clustering solutions on resampled data (80% participants, without replacement, 100 times). **B** Consistency of subtypes between K-means and HYDRA. The ARI and AMI values were estimated between clustering solutions derived using K-means and HYDRA. **C** Consistency across data modalities for HYDRA-based subtypes. The ARI and AMI values were estimated between clustering solutions derived using different clustering features. Bounds of the box represent the 1st (25%) and 3rd (75%) quartiles, the central line represents the median, the whiskers represent the values within 1.5 times of the interquartile range, the flier points represent outliers falling beyond the whiskers.

Qualitative confirmation from an independent dataset

We have shown that different modalities/features separately identify internally valid subtypes with similar clinical/cognitive profiles, and the concordance between subtypes derived from different modalities/features was low. Are these observations specific to the dataset we examined? We qualitatively replicated these findings using data from the REST-meta-MDD Project [54]. We separately identified subtypes using three different feature types: Conn, ReHo, or fALFF, and compared their available clinical measures. The HAMD measures were not significantly different between the ReHo-based subtypes ($p = 0.52$, Cohen's d = 0.05), or between the fALFF-based subtypes ($p = 0.76$, Cohen's d = 0.02). Although the HAMD measures were different between the Conn-based subtypes (Fig. 4A, $p = 0.020$, Cohen's d = 0.19), this comparison did not survive the correction for multiple comparisons (corrected for three feature types, FDR-corrected p = 0.061). For all feature types, the identified subtypes were not significantly different in their HAMA and illness duration (all p-values > 0.12). The subtypes were reproducible across data perturbations for each feature type (Fig. 4B), while the consistency of subtype

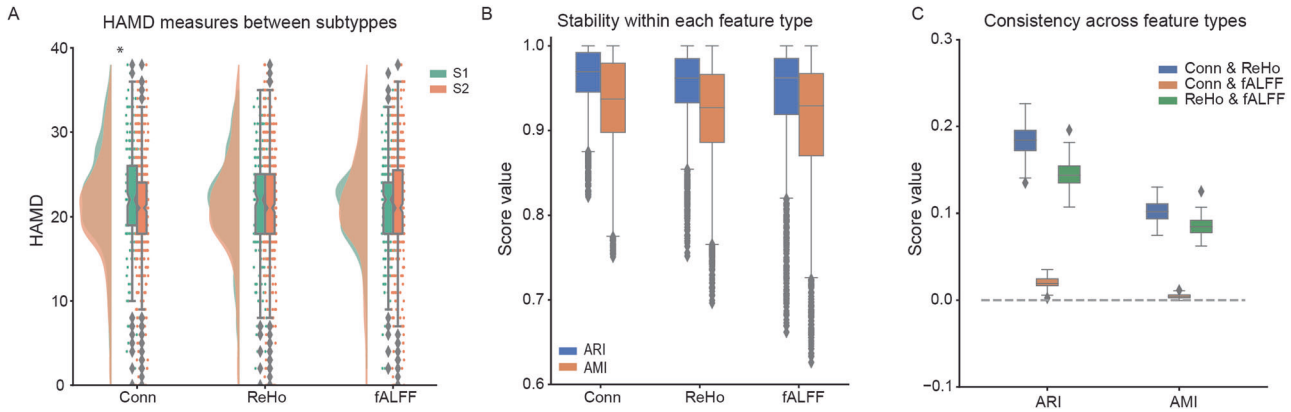

Fig. 4 Subtyping analysis on the external validation dataset. A Differences on the Hamilton Depression Rating Scale (HAMD) between subtypes identified using variants of features extracted from resting-state data. **B** Stability of subtypes identified with each type of feature. For each feature type, stability was estimated by calculating the adjusted rand index (ARI) and adjusted mutual information (AMI) between clustering solutions on resampled data (80% participants, without replacement, 100 times). **C** Consistency of subtypes identified using different types of features. The ARI and AMI values were estimated between clustering solutions derived using different clustering features. Bounds of the box represent the 1st (25%) and 3rd (75%) quartiles, the central line represents the median, the whiskers represent the values within 1.5 times of the interquartile range, the flier points represent outliers falling beyond the whiskers. *$p < 0.05$.

assignments between different feature types was low (ARI/AMI < 0.2, Fig. 4C). We also conducted the classification analysis on discriminating the two subtypes from healthy controls, which indicated that subtyping did not improve classification performance in this dataset (Figure S5). Overall, these results from this completely independent dataset with a different diagnostic group were consistent with those obtained from our main dataset.

DISCUSSION

We conducted subtyping analyses based on different neuroimaging features extracted from a sample of heterogeneous cases diagnosed with PTSD and/or TBI and healthy controls. We identified internally validated subtypes on the cases (i.e., combination of PTSD, TBI, and PTSD&TBI) with good separation and compactness from structural, task-based, and resting-state features. Although neurobiologically distinct, the clinical and/or cognitive profiles of these subtypes were not significantly different after correcting for multiple comparisons, and the subtyping did not improve case-control classification performance. Importantly, the composition of cases within the subtypes identified using one neuroimaging modality was very different from those identified by another imaging modality. Moreover, different feature types extracted from the same imaging modality led to distinct subtypes. We observed similar patterns of results from another independent dataset. Collectively, our results suggest that while internally validated and neurobiologically distinct subtypes could be generated using a given neuroimaging modality, the choice of clustering input may lead to high variability in neuroimaging-based subtyping.

Significance and stability of the clustering solutions are important properties to be tested for internal validation of subtyping analyses. Significance test assesses if the identified subtypes can be observed even in unstructured data (i.e., no underlying clusters). Stability test assesses if small perturbations of the data will largely change the cluster solution. In our analyses, all the identified subtypes met these internal validation criteria, and the subtypes were consistent across clustering methods, suggesting that these subtypes do capture the heterogeneity of the neuroimaging data. However, these significant and stable modality-specific subtypes showed very low consistency across modalities or feature types, such that pairs of individuals in the same subtype in one modality are not necessarily in the same subtype in another modality. Furthermore, different feature types extracted from the same neuroimaging modality could lead to distinct subtypes, which makes the identified subtypes highly specific to the analytic strategy, and thus less reproducible. Similar variabilities were observed in previous studies, where the authors found that activations from different tasks [60], or functional connectivity from different networks [49], resulted in variants of subtypes in healthy controls and patients. Together, these results highlight the importance of considering the variability in neuroimaging-based subtyping analysis within the context of reproducibility.

Clinical utility is a key aspect of subtyping [12, 61]. Distinct subtyping solutions based on different features can be meaningful if they capture different clinical aspects of data. On the other hand, even for highly reproducible subtypes, their clinical utilities are limited if they do not guide or inform clinical practices. In our analyses, we observed limited differences in clinical and cognitive profiles between subtypes identified using either data modality. Although there were some between-subtypes differences in the examined measures under liberal statistical thresholds, the effect sizes were very small, which has limited potential impacts on clinical practices. Similarly, we did not observe advantages of subtyping to improve neuroimaging-based diagnostic accuracy in distinguishing cases from controls. The limited clinical differences between subtypes are not uncommon in the literature. The reported associations between neuroimaging-based subtypes and clinical profiles were usually small to moderate, and even weaker on replication datasets [49]. These effect sizes may also be inflated by publication bias [62], as studies do not replicate previous results or do not find clinically meaningful subtypes are harder to be published. As reviewed recently, only a small portion of biologically identified subtypes in the literature showed potential for their clinical utility [12]. With the increasing availability of different data modalities, we have more analytical flexibility in our subtyping strategies. This large variability of subtyping analysis might increase the false positive rates in reporting clinical differences between identified subtypes, which requires attention and caution in interpreting subtyping results.

There are several potential explanations for the observed subtyping variability and limited between-subtypes clinical differences. First, the low reliability of neuroimaging-based features may increase variability across modalities. Although the anatomical measures are reliable across scanning sessions, the test-retest reliabilities of functional connectivity and task-based activations are much lower [63, 64]. Second, the clinical and

cognitive measures could be noisy. Although clinical measures from structured interview with clinicians (e.g., CAPS for PTSD) have strong interrater and test-retest reliability, the reliability of both self-report and cognitive task measures is imperfect [65, 66], which will likely impact the detection of brain-behavior associations [67]. Third, we only examined a limited number of clinical/cognitive measures, most of which were summary scores. It is possible that the subtypes differed in some specific clinical/cognitive domains/subdomains that were not captured by the measures we used. Fourth, although the subtypes exhibited similar cross-sectional clinical/cognitive profiles, they may diverge longitudinally. The subtypes might also be different in their treatment responses (e.g., refs. [21, 68]), which could not be captured in cross-sectional studies. Longitudinal studies with a wide range of symptom trajectories (e.g., [69, 70]) are needed to advance our understanding of biologically identified subtypes. Fifth, clinically relevant signals might be obscured by other factors such as demographic differences and/or structured noise. These factors may be difficult to identify and be removed, thus limit the ability to delineate clinical-relevant heterogeneity in clustering analysis. Sixth, the effect sizes of between-subtypes clinical differences can be small, necessitating larger samples from a more diverse source, such as different trauma types of PTSD, to detect the biological-clinical associations. Although the sample sizes we used here are comparable or larger than most neuroimaging-based subtyping studies, recent study has demonstrated that thousands of individuals may be required in some cases to estimate reproducible brain-behavior associations [71]. The objective of this study is not to dispute the existence of neuro-subtypes; rather, we aimed to demonstrate the difficulty in identifying clinically meaningful subtypes in two representative datasets, and the significant variability in neuroimaging-based subtyping analysis.

To move forward, several aspects can be considered for identifying biotypes with potential utility. First, hybrid analytic methods, which integrate a prior hypothesis (i.e., theory-driven) with data-driven clustering methods, may provide more insights. Considering the significant impact of clustering features on the subtyping results, it would be beneficial to use theoretical hypotheses to restrict the exploration spaces. For example, PTSD is characterized by abnormal structural volumes [72, 73], functional connectivity [74, 75], and task-based activations [76, 77] in the literature. This knowledge can potentially be used to determine clustering features. Second, exact replication studies are needed for testing the generalizability of reported subtypes. Although many different subtypes have been identified across studies, attempts to replicate these results are relatively sparse. A few replication studies did not reproduce the reported subtypes (e.g., refs. [26–28]). However, since these were usually "conceptual nonexact replications", it is difficult to determine the reasons for the failures. Since slight differences in sample characteristics, feature measures, or methodologies can significantly impact subtyping solutions, exact replications of promising subtypes are crucial for their clinical utility. Third, new methods with stringent validations may facilitate the identification of subtypes. On the one hand, new approaches such as individualized brain mapping [78] and normative modeling [79] can provide more reliable and informative features for clustering algorithms. On the other hand, integrating features from different imaging modalities may better capture the heterogeneity in patients [80]. Advanced multimodal fusion algorithms from the machine learning field [81] may offer advantages over traditional methods.

Several limitations of the present study should be considered. First, only two commonly used clustering methods were used to define subtypes. There are many other clustering methods in the field, such as the 3 C Algorithm [16] and canonical correlation analysis (CCA)-based approaches [21, 82] that combine both biological and clinical measures for identifying subtypes. Whether the observations in this study apply to these other methods should be tested in future research. In CCA-based approaches, since biological measures and clinical measures are integrated to derive clustering features, it is essential to be cautious when comparing the clinical profiles between identified subtypes; only measures not used in the clustering analysis should be compared to avoid bias. Second, raw features extracted from each modality were used for subtyping. Although this strategy was widely applied in the literature, it might not be the optimal strategy for feature extraction. Previous studies have shown the potential of feature abnormality [20] or normative modeling [79] in the subtyping procedure. It is unclear whether these feature extraction methods would improve the subtyping solutions by decreasing variability and increasing clinical utility. Third, the participants included in the main dataset were war-zone-exposed veterans, which may not represent the general population. At the same time, we observed similar variability on an independent civilian MDD dataset, suggesting the generalizability of the primary results. Future studies including different populations should be conducted to confirm the results.

In summary, our results highlight that even if internally validated subtypes are identified, the clinical and/or cognitive profiles of these subtypes may largely overlap, limiting their clinical utility. Therefore, caution is warranted in interpreting neuroimaging-based subtyping results. Additionally, analysis choices, such as clustering features, significantly impact the identification of subtypes, which should be carefully considered in future studies. Neuroimaging-based subtyping may be a promising approach to advance precision psychiatry, but rigorous validation of the subtypes grounded in clinical goals is needed.

DATA AVAILABILITY

The main dataset is available from CRM (Charles.marmar@nyulangone.org) upon reasonable request. The data are not publicly available due to a lack of informed consent from the participants and ethical approval for public data sharing. The REST-meta-MDD dataset is available at https://doi.org/10.57760/sciencedb.o00115.00013.

CODE AVAILABILITY

Codes for the main analyses are available at https://github.com/zhenfu-wen01/subtype_variability.

REFERENCES

1. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry. 2010;167:748–51.
2. Bryant RA, Galatzer-Levy I, Hadzi-Pavlovic D. The heterogeneity of posttraumatic stress disorder in DSM-5. JAMA Psychiatry. 2023;80:189–91.
3. Galatzer-Levy IR, Bryant RA. 636,120 ways to have posttraumatic stress disorder. Perspect Psychol Sci. 2013;8:651–62.
4. Pitman RK, Rasmusson AM, Koenen KC, Shin LM, Orr SP, Gilbertson MW, et al. Biological studies of post-traumatic stress disorder. Nat Rev Neurosci. 2012;13:769–87.
5. Hinojosa CA, George GC, Ben-Zion Z Neuroimaging of posttraumatic stress disorder in adults and youth: progress over the last decade on three leading questions of the field. Mol Psychiatry. 2024;29:1–22.
6. Shalev A, Liberzon I, Marmar C. Post-Traumatic Stress Disorder. N Engl J Med. 2017;376:2459–69.
7. Gong Q, He Y. Depression, Neuroimaging and Connectomics: A Selective Overview. Biol Psychiatry. 2015;77:223–35.
8. Chai Y, Sheline YI, Oathes DJ, Balderston NL, Rao H, Yu M. Functional connectomics in depression: insights into therapies. Trends Cogn Sci. 2023;27:814–32.
9. Shin LM, Liberzon I. The Neurocircuitry of Fear, Stress, and Anxiety Disorders. Neuropsychopharmacology. 2010;35:169–91.
10. Craske MG, Stein MB, Eley TC, Milad MR, Holmes A, Rapee RM, et al. Anxiety disorders. Nat Rev Dis Primer. 2017;3:1–19.
11. Feczko E, Miranda-Dominguez O, Marr M, Graham AM, Nigg JT, Fair DA. The heterogeneity problem: approaches to identify psychiatric subtypes. Trends Cogn Sci. 2019;23:584–601.

12. Brucar LR, Feczko E, Fair DA, Zilverstand A. Current approaches in computational psychiatry for the data-driven identification of brain-based subtypes. Biol Psychiatry. 2023;93:704–16.
13. Lanius RA, Vermetten E, Loewenstein RJ, Brand B, Schmahl C, Bremner JD, et al. Emotion modulation in PTSD: clinical and neurobiological evidence for a dissociative subtype. Am J Psychiatry. 2010;167:640–7.
14. van Loo HM, de Jonge P, Romeijn J-W, Kessler RC, Schoevers RA. Data-driven subtypes of major depressive disorder: a systematic review. BMC Med. 2012;10:156.
15. Fine NB, Ben-Zion Z, Biran I, Hendler T. Neuroscientific account of Guilt- and Shame-Driven PTSD phenotypes. Eur J Psychotraumatology. 2023;14:2202060.
16. Ben-Zion Z, Zeevi Y, Keynan NJ, Admon R, Kozlovski T, Sharon H, et al. Multidomain potential biomarkers for post-traumatic stress disorder (PTSD) severity in recent trauma survivors. Transl Psychiatry. 2020;10:1–11.
17. Hong S-J, Vogelstein JT, Gozzi A, Bernhardt BC, Yeo BTT, Milham MP, et al. Toward neurosubtypes in autism. Biol Psychiatry. 2020;88:111–28.
18. Marquand AF, Wolfers T, Mennes M, Buitelaar J, Beckmann CF. Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. Biol Psychiatry Cogn Neurosci Neuroimaging. 2016;1:433–47.
19. Etkin A, Maron-Katz A, Wu W, Fonzo GA, Huemer J, Vértes PE, et al. Using fMRI connectivity to define a treatment-resistant form of post-traumatic stress disorder. Sci Transl Med. 2019;11:eaal3236.
20. Maron-Katz A, Zhang Y, Narayan M, Wu W, Toll RT, Naparstek S, et al. Individual patterns of abnormality in resting-state functional connectivity reveal two data-driven PTSD subgroups. Am J Psychiatry. 2020;177:244–53.
21. Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med. 2017;23:28–38.
22. Liang S, Deng W, Li X, Greenshaw AJ, Wang Q, Li M, et al. Biotypes of major depressive disorder: neuroimaging evidence from resting-state default mode network patterns. NeuroImage Clin. 2020;28:102514.
23. Honnorat N, Dong A, Meisenzahl-Lechner E, Koutsouleris N, Davatzikos C. Neuroanatomical heterogeneity of schizophrenia revealed by semi-supervised machine learning methods. Schizophr Res. 2019;214:43–50.
24. Jiang Y, Wang J, Zhou E, Palaniyappan L, Luo C, Ji G, et al. Neuroimaging biomarkers define neurophysiological subtypes with distinct trajectories in schizophrenia. Nat Ment Health. 2023;1:186–99.
25. Stevens JS, Harnett NG, Lebois LAM, van Rooij SJH, Ely TD, Roeckner A, et al. Brain-based biotypes of psychiatric vulnerability in the acute aftermath of trauma. Am J Psychiatry. 2021;178:1037–49.
26. Dinga R, Schmaal L, Penninx BWJH, van Tol MJ, Veltman DJ, van Velzen L, et al. Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). NeuroImage Clin. 2019;22:101796.
27. Esterman M, Stumps A, Jagger-Rickels A, Rothlein D, DeGutis J, Fortenbaugh F, et al. Evaluating the evidence for a neuroimaging subtype of posttraumatic stress disorder. Sci Transl Med. 2020;12:eaaz9343.
28. Ben-Zion Z, Spiller TR, Keynan JN, Admon R, Levy I, Liberzon I, et al. Evaluating the evidence for brain-based biotypes of psychiatric vulnerability in the acute aftermath of trauma. Am J Psychiatry. 2023;180:146–54.
29. Liu Y, Hayes DN, Nobel A, Marron JS. Statistical significance of clustering for high-dimension, low-sample size data. J Am Stat Assoc. 2008;103:1281–93.
30. Stein MB, Bomyea J. Replicating predictive cluster-based imaging biotypes after trauma: a bridge too far? Am J Psychiatry. 2023;180:114–6.
31. Zang Y, Jiang T, Lu Y, He Y, Tian L. Regional homogeneity approach to fMRI data analysis. NeuroImage. 2004;22:394–400.
32. Suárez LE, Markello RD, Betzel RF, Misic B Linking Structure and Function in Macroscale Brain Networks. Trends Cogn Sci. 2020. 24 February 2020. https://doi.org/10.1016/j.tics.2020.01.008.
33. Yan C-G, Yang Z, Colcombe SJ, Zuo X-N, Milham MP. Concordance among indices of intrinsic brain function: Insights from inter-individual variation and temporal dynamics. Sci Bull. 2017;62:1572–84.
34. Etkin A, Egner T, Peraza DM, Kandel ER, Hirsch J. Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. Neuron. 2006;51:871–82.
35. Fonzo GA, Etkin A, Zhang Y, Wu W, Cooper C, Chin-Fatt C, et al. Brain regulation of emotional conflict predicts antidepressant treatment response for depression. Nat Hum Behav. 2019;3:1319–31.
36. Gaser C, Dahnke R, Thompson PM, Kurth F, Luders E, Initiative ADN. CAT – a computational anatomy toolbox for the analysis of structural MRI data. 2022:2022.06.11.495736.
37. Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods. 2019;16:111–6.
38. Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo X-N, Holmes AJ, et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb Cortex. 2018;28:3095–114.
39. Tian Y, Margulies DS, Breakspear M, Zalesky A. Topographic organization of the human subcortex unveiled with functional connectivity gradients. Nat Neurosci. 2020;23:1421–32.
40. King M, Hernandez-Castillo CR, Poldrack RA, Ivry RB, Diedrichsen J. Functional boundaries in the human cerebellum revealed by a multi-domain task battery. Nat Neurosci. 2019;22:1371–8.
41. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. NeuroImage. 2017;161:149–70.
42. Yu M, Linn KA, Cook PA, Phillips ML, McInnis M, Fava M, et al. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum Brain Mapp. 2018;39:4213–27.
43. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. NeuroImage. 2018;167:104–20.
44. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
45. Schubert E. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. ACM SIGKDD Explor Newsl. 2023;25:36–42.
46. Chen YT, Witten DM. Selective inference for k-means clustering. J Mach Learn Res. 2023;24:152–1.
47. Hivert B, Agniel D, Thiébaut R, Hejblum BP. Post-clustering difference testing: Valid inference and practical considerations with applications to ecological and biological data. Comput Stat Data Anal. 2024;193:107916.
48. Silverstein SM, Berten S, Olson P, Paul R, Williams LM, Cooper N, et al. Development and validation of a World-Wide-Web-based neurocognitive assessment battery: WebNeuro. Behav Res Methods. 2007;39:940–9.
49. Urchs SG, Tam A, Orban P, Moreau C, Benhajali Y, Nguyen HD, et al. Functional connectivity subtypes associate robustly with ASD diagnosis. eLife. 2022;11:e56257.
50. Murphy K, Fox MD. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. NeuroImage. 2017;154:169–73.
51. Saad ZS, Gotts SJ, Murphy K, Chen G, Jo HJ, Martin A, et al. Trouble at Rest: How Correlation Patterns and Group Differences Become Distorted After Global Signal Regression. Brain Connect. 2012;2:25–32.
52. Li J, Kong R, Liégeois R, Orban C, Tan Y, Sun N, et al. Global signal regression strengthens association between resting-state functional connectivity and behavior. NeuroImage. 2019;196:126–41.
53. Liu TT, Nalci A, Falahpour M. The global signal in fMRI: Nuisance or Information? NeuroImage. 2017;150:213–29.
54. Chen X, Lu B, Li H-X, Li X-Y, Wang Y-W, Castellanos FX, et al. The DIRECT consortium and the REST-meta-MDD project: towards neuroimaging biomarkers of major depressive disorder. Psychoradiology. 2022;2:32–42.
55. Yan C-G, Chen X, Li L, Castellanos FX, Bai T-J, Bo Q-J, et al. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. Proc Natl Acad Sci. 2019;116:9078–83.
56. Zou Q-H, Zhu C-Z, Yang Y, Zuo X-N, Long X-Y, Cao Q-J, et al. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. J Neurosci Methods. 2008;172:137–41.
57. Hivert B, Agniel D, Thiébaut R, Hejblum BP. Post-clustering difference testing: valid inference and practical considerations. 2022.
58. Varol E, Sotiras A, Davatzikos C. HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. NeuroImage. 2017;145:346–64.
59. Wen J, Fu CHY, Tosun D, Veturi Y, Yang Z, Abdulkadir A, et al. Characterizing heterogeneity in neuroimaging, cognition, clinical symptoms, and genetics among patients with late-life depression. JAMA Psychiatry. 2022;79:464–74.
60. Hawco C, Dickie EW, Jacobs G, Daskalakis ZJ, Voineskos AN. Moving beyond the mean: Subgroups and dimensions of brain activity and cognitive performance across domains. NeuroImage. 2021;231:117823.
61. Agelink van Rentergem JA, Deserno MK, Geurts HM. Validation strategies for subtypes in psychiatry: a systematic review of research on autism spectrum disorder. Clin Psychol Rev. 2021;87:102033.
62. Ioannidis JPA, Munafò MR, Fusar-Poli P, Nosek BA, David SP. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends Cogn Sci. 2014;18:235–41.
63. Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, et al. What is the test-retest reliability of common task-functional mri measures? New empirical evidence and a meta-analysis. Psychol Sci. 2020;31:792–806.
64. Noble S, Scheinost D, Constable RT. A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. NeuroImage. 2019;203:116157.

65. Enkavi AZ, Eisenberg IW, Bissett PG, Mazza GL, MacKinnon DP, Marsch LA, et al. Large-scale analysis of test–retest reliabilities of self-regulation measures. Proc Natl Acad Sci. 2019;116:5472–7.
66. Enkavi AZ, Poldrack RA. Implications of the lacking relationship between cognitive task and self report measures for psychiatry. Biol Psychiatry Cogn Neurosci Neuroimaging. 2021;6:670–2.
67. Gell M, Eickhoff SB, Omidvarnia A, Küppers V, Patil KR, Satterthwaite TD, et al. The burden of reliability: how measurement noise limits brain-behaviour predictions. 2023:2023.02.09.527898.
68. Zhang Y, Wu W, Toll RT, Naparstek S, Maron-Katz A, Watts M, et al. Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography. Nat Biomed Eng. 2021;5:309–23.
69. McLean SA, Ressler K, Koenen KC, Neylan T, Germine L, Jovanovic T, et al. The AURORA Study: a longitudinal, multimodal library of brain biology and function after traumatic stress exposure. Mol Psychiatry. 2020;25:283–96.
70. Ben-Zion Z, Fine NB, Keynan NJ, Admon R, Halpern P, Liberzon I, et al. Neurobehavioral moderators of post-traumatic stress disorder (PTSD) trajectories: study protocol of a prospective MRI study of recent trauma survivors. Eur J Psychotraumatol. 2019;10:1683941.
71. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, et al. Reproducible brain-wide association studies require thousands of individuals. Nature. 2022;603:654–60.
72. Ben-Zion Z, Korem N, Fine NB, Katz S, Siddhanta M, Funaro MC, et al. Structural Neuroimaging of Hippocampus and Amygdala Subregions in Posttraumatic Stress Disorder: A Scoping Review. Biol Psychiatry Glob Open Sci. 2024;4:120–34.
73. Wang X, Xie H, Chen T, Cotton AS, Salminen LE, Logue MW, et al. Cortical volume abnormalities in posttraumatic stress disorder: an ENIGMA-psychiatric genomics consortium PTSD workgroup mega-analysis. Mol Psychiatry. 2021;26:4331–43.
74. Ross MC, Cisler JM. Altered large-scale functional brain organization in post-traumatic stress disorder: A comprehensive review of univariate and network-level neurocircuitry models of PTSD. NeuroImage Clin. 2020;27:102319.
75. Bao W, Gao Y, Cao L, Li H, Liu J, Liang K, et al. Alterations in large-scale functional networks in adult posttraumatic stress disorder: a systematic review and meta-analysis of resting-state functional connectivity studies. Neurosci Biobehav Rev. 2021;131:1027–36.
76. Kredlow MA, Fenster RJ, Laurent ES, Ressler KJ, Phelps EA. Prefrontal cortex, amygdala, and threat processing: implications for PTSD. Neuropsychopharmacology. 2021;47:247–59.
77. Joshi SA, Duval ER, Kubat B, Liberzon I. A review of hippocampal activation in post-traumatic stress disorder. Psychophysiology. 2020;57:e13357.
78. Gordon EM, Laumann TO, Gilmore AW, Newbold DJ, Greene DJ, Berg JJ, et al. Precision functional mapping of individual human brains. Neuron. 2017;95:791–807.
79. Rutherford S, Kia SM, Wolfers T, Fraza C, Zabihi M, Dinga R, et al. The normative modeling framework for computational psychiatry. Nat Protoc. 2022;17:1711–34.
80. Calhoun VD, Sui J. Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link(s) in Complex Mental Illness. Biol Psychiatry Cogn Neurosci Neuroimaging. 2016;1:230–44.
81. Lahat D, Adali T, Jutten C. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. Proc IEEE. 2015;103:1449–77.
82. Wang H-T, Smallwood J, Mourao-Miranda J, Xia CH, Satterthwaite TD, Bassett DS, et al. Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. NeuroImage. 2020;216:116745.

ACKNOWLEDGEMENTS

This study was supported by grants from the Steven A. and Alexandra M. Cohen Foundation, Inc., Cohen Veterans Bioscience, Inc., National Institute of Mental Health grants R01MH123736, R01MH125198 and R33MH111907.

AUTHOR CONTRIBUTIONS

Conceptualization: ZW, MRM, and CRM. Methodology: ZW, MZH, CES, EML, MRM, CRM. Data collection: DA-A, AE, CRM. Formal analysis: ZW, MZH. Supervision: MRM, CRM. Writing-original draft: ZW, MRM. Writing-review & editing: ZW, MZH, CES, EML, AE, MRM, CRM. Funding acquisition: AE, MRM, CRM.

COMPETING INTERESTS

CRM has served on advisory boards of Receptor Life Sciences, Otsuka Pharmaceuticals and Roche Products Limited and has received support from the National Institute on Alcohol Abuse and Alcoholism, National Institute of Mental Health, Department of Defense-CDMRP*US Army Research Office*DARPA, Bank of America Foundation, Brockman Foundation, Cohen Veterans Bioscience, Cohen Veterans Network, McCormick Foundation, Home Depot Foundation, New York City Council, New York State Health, Mother Cabrini Foundation, Tilray Pharmaceuticals, Ananda Scientific and GrayMatters Health. AE reports salary and equity from Alto Neuroscience. The other authors report no financial relationships with commercial interests.

ETHICAL APPROVAL AND CONSENT TO PARTICIPATE

All study procedures were performed in accordance with approved guidelines and relevant regulations. The study was reviewed and approved by the Institutional Review Boards (IRBs) of the NYU Grossman School of Medicine (IRB no. i12-03926) and Stanford University (IRB no. 5136). All participants provided written informed consent.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41380-024-02807-y.

Correspondence and requests for materials should be addressed to Mohammed R. Milad or Charles R. Marmar.