

## ARTICLE OPEN



# Characterizing replicability in the clustering structure of brain morphology in autism, attention-deficit/hyperactivity disorder, and obsessive compulsive disorder

Younes Sadat-Nejad<sup>1,2</sup>, Marlee M. Vandewouw<sup>1,3</sup>, Jessica Brian<sup>1</sup>, Jennifer Crosbie<sup>2,4</sup>, Russell J. Schachar<sup>2,4</sup>, Alana Iaboni<sup>1</sup>, Elizabeth Kelley<sup>5</sup>, Jessica Jones<sup>5</sup>, Margot J. Taylor<sup>2,4</sup>, Muhammad Ayub<sup>6</sup>, Robert Nicolson<sup>7</sup>, Bilal Syed<sup>4</sup>, Christopher Hammill<sup>4</sup>, Stelios Georgiades<sup>8</sup>, Paul D. Arnold<sup>9</sup>, Jason P. Lerch<sup>2,10</sup>, Evdokia Anagnostou<sup>1,2,11</sup> and Azadeh Kushki<sup>1,2</sup>✉

© The Author(s) 2025

In neurodevelopmental research, within-diagnosis heterogeneity and across-diagnosis overlap necessitate a shift from case-control designs to data-driven clustering approaches. However, our understanding of the replicability of these clustering structures across independent datasets remains limited. Our objective was to examine the replicability of clustering structure in measures of brain morphology in neurodiverse children across two independent datasets, namely the Province of Ontario Neurodevelopmental Disorder (POND) Network and the Healthy Brain Network (HBN). POND and HBN data were collected across various institutions in Ontario, Canada, and New York, United States, respectively. Participants were 5–19 years old and had diagnoses of autism, attention deficit/hyperactivity disorder (ADHD), obsessive compulsive disorder (OCD), or were neurotypical. We used measures of cortical volume, surface area, cortical thickness, and subgroup volume from structural MRI data. Principal component analysis (PCA) and clustering were used to examine the replicability of clustering structures across the datasets. Correlations among principle components, measures of clusterability, and alignment between the four brain measures as well as male/female subsets were examined. Brain-behaviour associations were examined using univariate and multivariate approaches. The POND dataset included 747 participants with (autism  $n = 312$ , ADHD  $n = 220$ , OCD  $n = 70$ , neurotypical  $n = 145$ ). The HBN dataset included 582 participants (autism  $n = 60$ , ADHD  $n = 445$ , OCD  $n = 19$ , neurotypical  $n = 58$ ). Our results showed significant between-dataset correlations in 82.1% of the principal components derived from brain measures. A two-cluster structure was replicated across datasets, brain measures, and the female/male subsets, however the participant composition of clusters were only aligned between cortical volume and surface area, and cortical thickness and subcortical volume. Regional effect sizes for between-cluster differences were highly correlated across datasets ( $\beta = 0.92 \pm 0.01$ ,  $p < 0.0001$ ; adjusted  $R^2 = 0.93$ ). Data-driven clusters did not align with diagnostic labels across datasets. Brain-behaviour associations were only replicated for male subsets and subcortical volume using multivariate analysis. We found evidence of replicability of the clustering structure across two independent datasets; however, caution must be exercised in integrating multiple measures in clustering and interpretation of brain-behaviour associations.

*Translational Psychiatry* (2025)15:333; <https://doi.org/10.1038/s41398-025-03540-y>

## INTRODUCTION

Autism spectrum disorder (ASD; autism), attention-deficit/hyperactivity disorder (ADHD), and obsessive-compulsive disorder (OCD) are behaviourally-defined neurodevelopmental conditions [1–3] with significant variability and overlap in their neurobiology and phenotypic presentation [4, 5]. To characterize the variability within and across these conditions, a growing body of research has focused on data-driven approaches, including clustering [6–9], to discover transdiagnostic groups of individuals who share similar neurobiological [7, 9–11] or phenotypic features [12]. These studies have consistently found a misalignment between data-

driven subgroups and existing diagnostic labels [6, 12]; however, significant variability exists across these studies in the neurobiological features and analytical approaches used in clustering. For example, different measures of brain morphology [12, 13] and function [10] have been used along with a range of clustering approaches including hierarchical, spectral, multi-view, or regression clustering [12, 13].

In addition to differences in data modalities and analytical approaches, significant variability is also found in data acquisition methods (e.g., scanners, scanning parameters, motion), imaging pipelines [14–16] (quality control method, denoising/correction

<sup>1</sup>Holland Bloorview Kids Rehabilitation Hospital, 150 Kilgour Rd, Toronto, ON M4G1R8, Canada. <sup>2</sup>University of Toronto, Toronto, Ontario, Canada. <sup>3</sup>Harvard Medical School, Boston, MA, USA. <sup>4</sup>The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>5</sup>Queen's University, Kingston, Ontario, Canada. <sup>6</sup>University College London, London, UK. <sup>7</sup>University of Western Ontario, London, Canada. <sup>8</sup>McMaster University, Hamilton, Canada. <sup>9</sup>University of Calgary, Calgary, Canada. <sup>10</sup>University of Oxford, Oxford, UK. <sup>11</sup>Department of Psychiatry, Icahn School of Medicine, Mount Sinai, NY, USA. ✉email: [akushki@hollandbloorview.ca](mailto:akushki@hollandbloorview.ca)

Received: 15 May 2024 Revised: 15 July 2025 Accepted: 12 August 2025

Published online: 30 August 2025

algorithms), and sample characteristics [17], including diagnoses and sociodemographic composition. Given this, it is not surprising that there is also significant variability in study findings. This includes differences in the suggested number of clusters (e.g., 2–8 cluster solutions [10, 12, 18]), neurobiological characteristics defining the subgroups (e.g. differences in cortical volume or subcortical [18], cortical thickness [12]), and the phenotypic presentation of the clusters (for example, cluster differences in social communication abilities [10, 18], language and attention [18], cognitive ability and hyperactivity [9]). The heterogeneity of findings in the existing literature has raised questions about the replicability of clustering results. In this context, replicability is defined as obtaining consistent findings across studies with the same research question [19]. Among the existing literature, only one study has investigated the replicability of data-driven subgroups across two independent datasets including children with diagnoses of neurodevelopmental conditions [9]. Using resting-state functional connectivity datasets from the Province of Ontario Neurodevelopmental Disorder (POND) Network and the Healthy Brain Network (HBN), this study found two clusters that differed in IQ, hyperactivity, and impulsivity, as well as patterns of segregation and integration within the brain's networks. These results provide encouraging preliminary evidence that the results of clustering based on these measures of brain function may be replicable in spite of differences in datasets. A critical gap still remains, however, in understanding replicability in clustering results based on measures of brain morphology. The present study addresses the gap by examining the issue of replicability in measures of brain morphology namely cortical thickness, surface area, and cortical and subcortical volume. To this end, we will examine 1) replicability of clustering across these measures, and 2) replicability across two independent datasets. Given the known sex-differences in neurodevelopmental conditions, our analysis was disaggregated by sex, allowing us to also examine replicability across the male and female subsets of each dataset.

## METHODS

### Participants

For this study, we used data from two independent datasets, namely, POND (export date May 22, 2023), and HBN (Release 10). Data from participants who were between 5–19 years of age, had a diagnosis of autism, ADHD, OCD, or who were neurotypical, and whose neuroimaging data passed quality control were selected for the current study. This resulted in data from 747 participants from POND (autism:  $n = 312$ , female = 22.4%, median age = 12.4 (5.53); ADHD:  $n = 220$ , 25.5% female, median age = 11.3 (4.08); OCD:  $n = 70$ , 40.0% female, median age = 11.8 (5.57); neurotypical:  $n = 145$ , 41.4% female, median age = 11.9 (5.16)), and 582 participants from HBN (autism:  $n = 60$ , 8.33% female, age = 12.31 (5.86); ADHD:  $n = 445$ , 31.2% female, 10.02 (4.83); OCD:  $n = 19$ , 52.6% female, age = 8.76 (5.09); neurotypical:  $n = 58$ , 41.4% female, age = 10.1 (5.24)). For POND, clinical diagnoses were supported by gold-standard assessments: the Autism Diagnostic Observation Schedule-2 (ADOS [20]) and the Autism Diagnostic Interview-Revised (ADI-R [20]) for autism, the Parent Interview for Children Symptoms for ADHD (PICS) for ADHD, and the Children's Yale-Brown Obsessive Compulsive Scale for OCD [21] (CY-BOCS). Children in the neurotypical group did not have a history of neurodevelopmental, psychiatric, or neurological diagnoses, were born after 35 weeks gestation, and had no first-degree relative with a neurodevelopmental condition. For HBN, a computerized web-based version of the Schedule for Affective Disorders and Schizophrenia—Children's version (KSADS [22]) was administered, which was reviewed alongside all study material by a clinical team to synthesize a consensus clinical diagnosis aligning with the DSM-5 [23]. Individuals with no diagnosis given were considered as neurotypical.

Both POND and HBN studies were approved by the respective institutions' research ethics board. Written informed consent and/or verbal assent (if written is not available) were obtained from the primary caregiver and/or participants as appropriate. The present study on secondary analysis of POND and HBN data was approved by the Holland Bloorview Research Ethics Board.

### Behavioural measures

Phenotypic measures available for both datasets included the Social Communication Questionnaire (SCQ) to quantify autism-like features [24], the Strength and Weakness for ADHD symptoms and Normal Behaviour Rating Scale (SWAN [25]) to measure inattention and hyperactivity symptoms, Toronto Obsessive Compulsive Scale (TOCS) to measure obsessive-compulsive traits, and full-scale intelligence quotient (FSIQ) measured by an age-appropriate IQ scale [26, 27]. The internalizing and externalizing measures of Child Behaviour Checklist (CBCL [28, 29]) were used to quantify internalizing and externalizing symptoms.

### Sociodemographics measures

In addition to age and sex, racial and ethnic identifications in both datasets were collected through self-reported or parent-reported questionnaires. For POND, racial categories were aligned with the standards set by the Canadian Institute for Health Information. These categories encompassed Black, East Asian, Indigenous, Latino, Middle Eastern, South Asian, Southeast Asian, White, and other. Participants with mixed racial backgrounds were coded in multiple categories. For HBN, racial categories followed the US Census guidelines, including American Indian or Alaskan Native, Asian, Black, Hispanic, Native Hawaiian or other Pacific Islander, White, 2 or more races, and other. Given the sample size, we consolidated race into two categories of minoritized and white for both datasets. For both datasets, household income was categorized as low (<\$74,999 CAD), medium (\$75,000 CAD to \$199,999 CAD), and high ( $\geq$ \$200,000 CAD). Education was defined based on the highest educational attainment of the primary caregiver, categorized as: Level 1 (non-completion of high school or high school diploma), Level 2 (associate degree or undergraduate degree), and Level 3 (graduate or professional degree).

### Imaging data

For both datasets, measures of cortical surface area, cortical thickness, cortical volume and subcortical volume were obtained from structural MRI (sMRI). For POND, the sMRI images were collected on Siemens MAGNETOM 3 T Trio and Prisma MRI scanners across three sites, namely, the Hospital for Sick Children (Toronto, Ontario; Trio:  $n = 233$ ; Prisma:  $n = 348$ ), Queen's University (Kingston, Ontario; Trio:  $n = 100$ ; Prisma:  $n = 43$ ), and Holland Bloorview Kids Rehabilitation Hospital (Toronto, Ontario; Prisma:  $n = 23$ ). For HBN, the data were collected using Siemens 3 T Trio and Prisma scanners from three institutions in the New York City area, namely the CitiGroup Cornell Brain Imaging Center (Prisma:  $n = 345$ ), Rutgers University (Trio:  $n = 202$ ), the City University of New York Advanced Science Research Center (Prisma:  $n = 28$ ), and a mobile site in Staten Island with a 1.5 Tesla Siemens Avanto ( $n = 7$ ).

To extract surface area, cortical thickness, and cortical volume, the CIVET pipeline (version 2.1.0) [30] was used. These measures were extracted for 76 regions based on the automated anatomical labeling atlas (AAL) [17, 31]. Non-uniformity image correction and stereotaxic registration to the Montreal Neurologic Institute (MNI ICBM) [32] template (non-linear 6th generation target) was then used. Masking, extraction and classification were used to separate and obtain gray matter, white matter, and cerebrospinal fluid volume. A surface diffusion kernel was applied [33], and regions were registered to the AAL atlas [34]. Cortical thickness was calculated based on the distance between two smooth surfaces [14] and gray matter and white matter surfaces was generated by tissue classification, and then surfaces were registered to the automated anatomical labelling (AAL) atlas [35]. Lastly, segmentation by use of multiple automatically generated templates (MAGeT) [33] was used to calculate volume of 95 subcortical structures from multiple starting atlases, including 5-atlas subcortical, cerebellum, amygdala, hippocampus-subfields, and striatum and thalamus subdivisions. The CIVET and MAGeT quality control (QC) pipelines were used, and participants were only included if they passed both QC pipelines. Details of the data filtering is provided in the eTable 1 in Supplement 1. For each dataset, separately for males and females, the brain measures were corrected for scanner effects using ComBat Harmonization [14]. For age correction, the best model fit among linear, quadratic, and cubic effects was used for each brain region [9].

### Analysis pipeline

Data and statistical analyses were performed using Python 3.8.0 and R 3.3.3. An overview of the analysis pipeline is depicted in eFigure 1 in Supplement 1. Given the sex-differences in neurobiology of

neurodevelopmental conditions [36–39], analyses were conducted independently on male and female subsets of each dataset.

To examine between-dataset similarities in the structure of the data, we used Principal Component Analysis (PCA) [40]. PCA is a multivariate approach that transforms the set of measurement variables into a new set of uncorrelated variables (principal components; PCs) that capture the largest variation in the data. The coefficients of the original variables, referred to as loadings, represent the strength of their contribution to the PCs. For this study, PCA was applied independently on surface area, cortical thickness, and cortical, and subcortical volume data. To examine similarities in the principal components across POND and HBNs, Pearson's correlation between POND and HBN loadings was computed. This was computed as the maximum correlation between a POND PC, and the corresponding PC on HBN, allowing for a window of 2 in cases where PC numbers were not aligned between the datasets.

To characterize the clustering structure of the datasets, we used the PCA-transformed data to compute participant similarity networks. These are matrices with entries corresponding to the similarity between pairs of participants (i.e., entry  $i, j$ , corresponds to the similarity between participants  $i$  and  $j$ ). Pairwise similarities were computed using the Gaussian transform of the cosine distance between vectors encoding brain measure values across all regions of the atlas. The cosine distance was selected as it provides a robust method for capturing structural associations in high-dimensional datasets [41]. With this pipeline, we generated four distinct similarity networks (cortical area, cortical thickness, cortical, and subcortical volume) separately for male and females in each dataset. These matrices were then clustered using spectral clustering [42].

### Statistical analyses

To examine the existence of clusters within each network, we employed three measures of clusterability: the gap statistic [43], silhouette coefficient [33], and Calinski-Harabasz [44]. Statistical significance of clustering patterns were determined using a permutation test comparing the three measures of clusterability between the datasets and 200 random networks. The random networks were generated using the same weight distribution as the original networks [45], preserving the degree and strength of the original networks [46]. Alignment among the constructed clusters and diagnostic labels, as well as clusters obtained from different brain measures was assessed using the adjusted rand score [47]. An adjusted rand score of one indicates full alignment between two sets, whereas a value of zero suggests no alignment.

Univariate and multivariate methods were employed to examine the associations among clusters and behavioural and brain measures. For univariate analysis, measures were compared among clusters using t-test for normally-distributed, continuous data, Mann-Whitney tests for non-normally-distributed, continuous data, and Chi-squared tests for categorical data. Family-wise correction was used for multiple comparisons and Cohen's effect size [42] was reported for statistically significant results. For multivariate analysis, we predicted cluster labels from phenotypic measures using a random forest classifier [48]. These phenotypic predictors included scores on the SCQ, SWAN (inattention and hyperactivity), and CBCL (internalizing and externalizing), as well as full-scale IQ, age, race/ethnicity, and household education level.

## RESULTS

### Participants

A total of 121 participants in POND and 923 participants in HBN failed either the CIVET and MAGet quality control (detailed in Supplementary Table 1). As the result, 747 participants from POND and 582 participants from HBN remained for the analyses. The demographic characteristics for the POND and HBN participants are shown in Table 1.

### PCA decomposition

The number of principal components needed to account for 75% of variance in the data across the measures and dataset ranged between 14 and 24 (eTable 2 in Supplement 1). The correlations between the loadings on the principal components of two datasets are shown in the eFigure 2 in Supplement 1. The loadings were significantly correlated between 82.1% of PCs. Of the

statistically significant correlations, 40.9% exceeded a correlation coefficient of 0.3. HBN females had the lowest percentage of significant correlations (eTable 3 in Supplement 1).

### Clustering composition of the data

Participant similarity matrices are visualized as network graphs in Fig. 1. As seen, two distinct groupings are evident across datasets, brain measures, and male and female subsets. To determine if clusters existed in the data, we used the gap statistic [43], comparing the within cluster dispersion of the data to that expected under the null distribution (no random permutation networks). The gap statistic was significantly larger for our data compared to the null distributions (random permutation networks) for surface area, cortical thickness, and cortical and subcortical volume for both males and females ( $p < 0.01$ , eTable 4 in Supplement 1). Silhouette and Calinski-Harabasz scores (eFigure 3 in Supplement 1) suggest that the optimal number of clusters is two for all measures and datasets.

### Clustering results

Across brain measures and datasets, there was very low alignment between diagnostic labels and data-driven groupings (adjusted rand scores  $< 0.02$ ; eTable 5 in Supplemental 1). Alignment among clusters constructed using different brain measures is shown in the eFigure 4 in Supplement 1. Across all datasets, clustering solutions were highly aligned for cortical volume and surface area (adjusted rand score 0.63–0.81), and moderately aligned for cortical thickness and subcortical volume (adjusted rand score 0.22–0.44). This finding was replicated across datasets and female/male subsets.

**Cluster differences in brain measures.** For both datasets, we computed the effect size for the differences in brain measures across clusters using Cohen's  $d$  (eFigure 5, eFigure 6 and eTable 6 in the Supplement 1). Figure 2 shows the association among these effect sizes between POND and HBN, as well as male and female subsets. Linear regression analysis revealed a significant association between cluster effect sizes for POND and HBN after controlling for measure and sex (intercept =  $0.09 \pm 0.02$ ,  $p < 0.0001$ ; beta =  $0.92 \pm 0.01$ ,  $p < 0.0001$ ; adjusted R-squared = 0.93). Similarly, a significant association was found for cluster effect sizes between males and females (intercept =  $-0.04 \pm 0.02$ ,  $p = 0.04$ ; beta =  $0.97 \pm 0.01$ ,  $p < 0.0001$ ; adjusted R-squared = 0.91). This suggests that brain signatures associated with the clusters are highly consistent between datasets and male/female subsets.

### Cluster associations with phenotypic measures

Univariate testing did not reveal any significant between-cluster-differences in age, race/ethnicity, family income and education, ethnicity, FSIQ, SCQ, SWAN, or CBCL scores (detailed statistics in the eTable 3 in Supplemental 1) across datasets or measures (Fig. 3).

The accuracy for multivariable prediction of cluster labels is presented in the eTable 7 in Supplemental 1. One-sample t-tests revealed that cluster labels were predicted with greater than chance accuracy for subcortical volume for males in both POND (accuracy =  $0.65 \pm 0.09$ ;  $p = 0.02$ ) and HBN (accuracy =  $0.61 \pm 0.05$ ;  $p = 0.01$ ). For those prediction tasks, feature importance values (calculated based on mean decrease in impurity [48]) are reported in supplemental eFigure 7. The differentiating features were highly consistent between POND and HBN and female and male subsets, with the highest importance attributed to age and the phenotypic measures (IQ, CBCL internalizing and external, SCQ, SWAN scores). The contribution of sociodemographic factors to prediction was significantly smaller.

**Table 1.** Demographic characteristics for the POND datasets.

		<b>Autism</b>	<b>ADHD</b>	<b>OCD</b>	<b>NT</b>	<b>All</b>
n	POND <sup>b</sup>	312	220	70	145	747
	HBN <sup>b</sup>	60	445	19	58	582
Age <sup>a</sup> (years)	POND <sup>b</sup>	12.26 (3.60)	11.63 (2.75)	13.16 (2.66)	12.15 (3.76)	12.14 (3.34)
	HBN <sup>b</sup>	10.02 (4.83)	12.31 (5.86)	11.43 (5.28)	8.76 (5.09)	10.12 (5.24)
Sex <sup>a</sup> No. (%) female	POND <sup>b</sup>	70 (22.4)	56 (25.5)	28 (40.0)	44 (43.1)	214 (28.6)
	HBN <sup>b</sup>	5 (8.3)	139 (31.2)	10 (52.6)	24 (42.8)	178 (41.4)
SCQ <sup>a</sup>	POND <sup>b</sup>	20.00 (10.00)	5.00 (7.00)	4.00 (6.00)	2.00 (2.00)	9.00 (16.00)
	HBN <sup>b</sup>	12.00 (7.50)	6.00 (5.00)	6.00 (9.00)	4.00 (4.00)	7.00 (6.00)
SWAN inattention	POND <sup>b</sup>	4.00 (5.00)	6.00 (4.00)	1.00 (3.25)	0.00 (0.00)	3.00 (7.00)
	HBN <sup>b</sup>	4.00 (6.00)	3.00 (6.00)	2.00 (3.50)	0.00 (0.25)	3.00 (6.00)
SWAN hyperactivity <sup>a</sup>	POND <sup>b</sup>	2.00 (4.00)	4.00 (6.00)	0.00 (2.00)	0.00 (0.00)	1.00 (5.00)
	HBN <sup>b</sup>	1.00 (4.00)	1.00 (4.00)	0.00 (2.00)	0.00 (0.00)	1.00 (3.00)
TOCS	POND <sup>b</sup>	−4.50 (34.00)	−23.00 (45.00)	20.00 (24.00)	−42.50 (48.00)	−12.50 (47.25)
Full-scale IQ	POND <sup>b</sup>	96.00 (29.50)	102.00 (18.50)	113.00 (23.50)	111.50 (13.75)	102.50 (23.0)
	HBN <sup>b</sup>	98.00 (25.00)	100.00 (24.0)	113.00 (14.00)	107.50 (19.75)	101.0 (25.0)
CBCL Internalizing <sup>a</sup>	POND <sup>b</sup>	65.00 (13.00)	63.00 (16.00)	68.00 (12.25)	48.00 (13.00)	61.00 (17.00)
	HBN <sup>b</sup>	61.00 (12.50)	59.00 (17.00)	63.50 (13.75)	49.00 (16.25)	58.00 (16.00)
CBCL Externalizing	POND <sup>b</sup>	58.00 (15.00)	61.00 (15.50)	53.00 (15.00)	43.50 (16.00)	55.00 (18.00)
	HBN <sup>b</sup>	59.00 (14.50)	58.00 (16.00)	57.50 (18.75)	49.00 (14.00)	57.00 (17.00)
Income low:med:high	POND <sup>b</sup>	44:51:24	34:33:25	0:6:4	16:28:26	94:118:79
	HBN <sup>b</sup>	16:17:10	109:125:120	3:4:8	10:14:24	138:160:162
Education (Level1:Level2:Level3) <sup>a</sup>	POND <sup>b</sup>	6:100:30	12:12:69	0:8:3	6:65:28	24:234:70
	HBN <sup>b</sup>	13:4:23	51:11:202	1:0:14	6:2:36	71:17:275
White No. (%)	POND <sup>b</sup>	152 (34.6)	143 (32.6)	34 (7.7)	110 (25.0)	439
	HBN <sup>b</sup>	23 (8.0)	219 (76.6)	10 (3.5)	34 (11.9)	286
Minoritized No. (%)	POND <sup>b</sup>	33 (60.0)	6 (10.9)	1 (1.8)	15 (27.3)	55
	HBN <sup>b</sup>	28 (12.8)	163 (74.4)	8 (3.6)	20 (9.1)	219

Values are reported as median (IQR) for continuous measures.

Reported values are median (interquartile range (IQR)). P values are for multiple comparisons (9 comparisons). P-values are results of t-test for normally distributed and Kruskal-Wallis test (for continuous non-normally distributed data). For sex, Chi-squared test was used.

<sup>a</sup>significant difference between datasets ( $p < 0.001$ ).

<sup>b</sup>significant difference between diagnostic groups ( $p < 0.001$ ).

## DISCUSSION

Our study characterized the replicability of the participant similarity networks constructed using surface area, cortical thickness, and cortical and subcortical volume, across the POND and HBN datasets, as well as male and female subsamples.

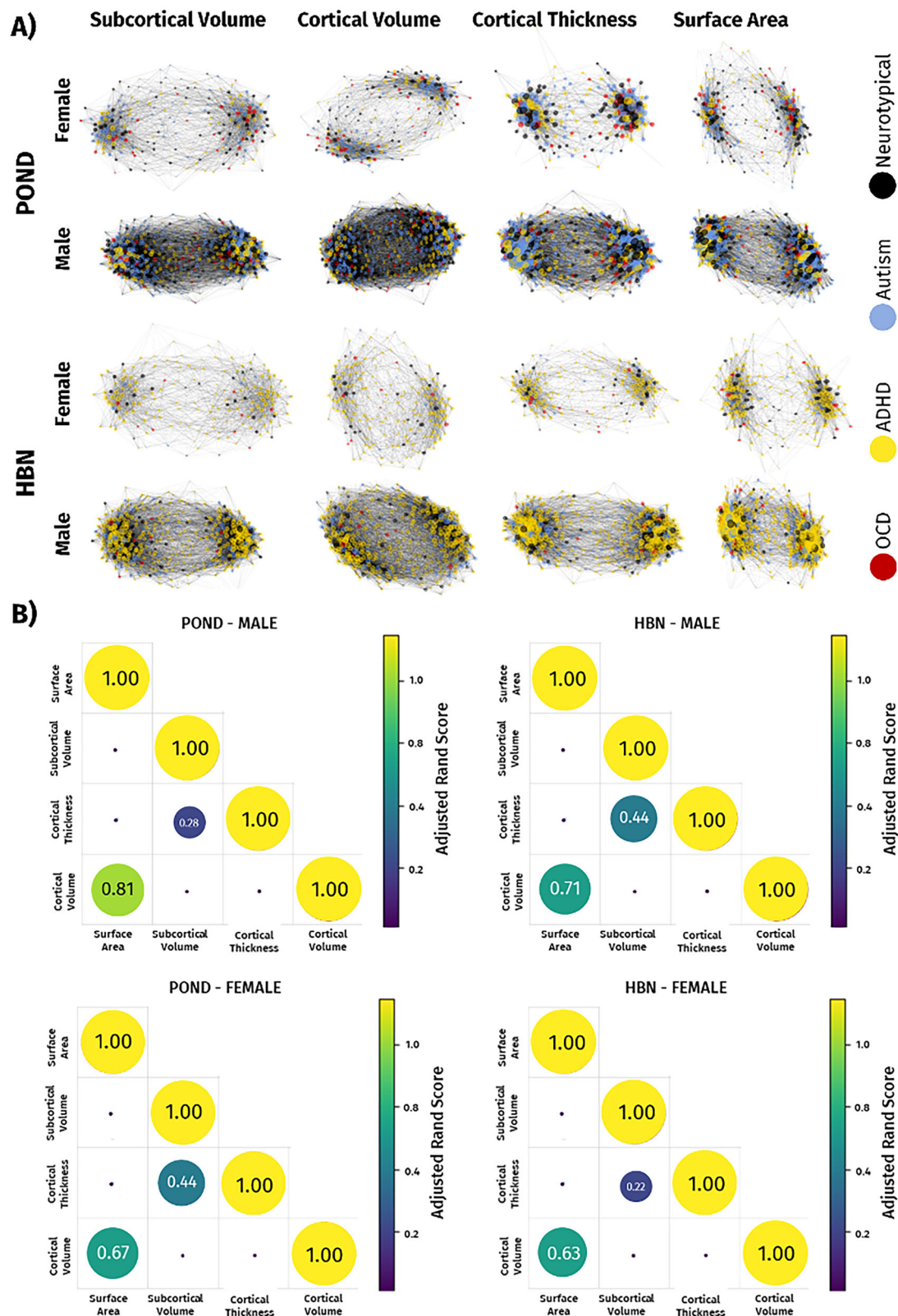
### Replicability across datasets

Despite significant differences in the POND and HBN datasets in demographic and phenotypic composition, our results revealed a high degree of consistency between the data structures for the two datasets. In particular, we found high between-datasets correlations among the principal components obtained using POND and HBN datasets, suggesting that data structures are similar in both datasets across the brain measures examined. The clustering structure was highly replicable across datasets, with our results revealing a 2-cluster composition across the four brain measures and the female/male subsets. This is at the lower end of previous literature findings where the number of reported clusters is highly variable, ranging from 2–10 [6, 9–11, 18, 49]. Larger number of clusters are likely to be found when multiple brain measures are combined, especially if these measures quantify potentially different biological mechanisms (for example, if two independent groups are found in each measure A and B, the combination of measures will result in four possible group combinations).

The brain signatures of the clusters were highly consistent across datasets with high correlations among regional effect sizes for between-cluster differences. Another finding that was replicated was that data-driven clusters were not aligned with diagnostic labels as indicated by the low Adjusted Rand Index scores (eTable 5 in Supplement), across datasets, brain measures, and female/male datasets. This finding is consistent with previous literature [6, 9, 10, 18, 49], further adding to the body of work highlighting the need for enhanced biologically-relevant precision in characterization of neurodevelopmental conditions, compared to our broad diagnostic categories.

In this study, we did not find statistically significant phenotypic differences between clusters through uni-variate analysis. However, multivariate analysis showed that cluster labels derived based on subcortical volume were predicted with greater than chance accuracy based on a combination of differences in age, IQ, internalizing, externalizing symptoms, autism features, and inattention and hyperactivity/impulsivity, across both POND and HBN. This finding suggests that neurobiological homogeneity may not align with single diagnostic domains of neurodevelopmental conditions, but instead, reflects differences in constellation of phenotypic features that are not specific to a single diagnosis category. The null finding of univariate phenotypic differences may also be due to statistical power as replicability in brain-behaviour associations may require very large sample sizes [50].



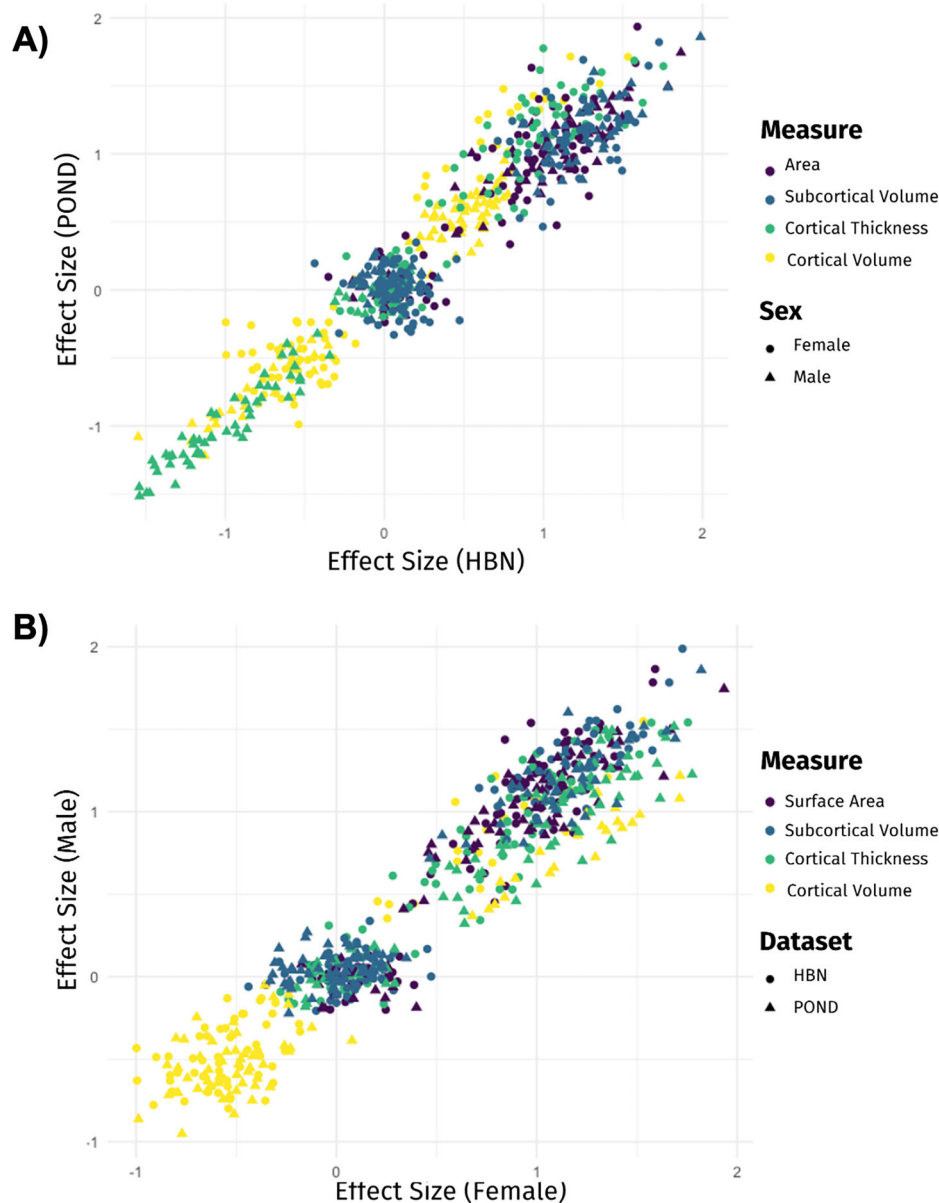


**Fig. 1** Replicability of clustering structure. **A** participant similarity networks, **(B)** adjusted rand score.

#### Replicability across brain measures

In addition to between-dataset differences, we examined the replicability of clustering structures across different brain measures within a dataset. The two-cluster solution was replicated across cortical area, cortical thickness, cortical volume, and

subcortical volume; However, the participant membership to the clusters was only partially replicated between subcortical volume and cortical thickness, and surface area and cortical volume, but not at all among other pairs of measures. The misalignment between cortical thickness and surface area is not surprising given



**Fig. 2** Association among between-cluster effect sizes computed. **A** POND and HBN, **(B)** male and female subsets.

that these features are suggested to be genetically distinct determinants of cortical structure [51]. Further, the finding of replicability between cortical volume and surface area is consistent with the suggestion that interindividual variation in gray matter volume is largely driven by differences in surface area rather than the cortical thickness [52]. The dissociation between cortical thickness and surface area is particularly important to studies of subgroup structure in neurodevelopmental conditions that integrate multiple measures of cortical morphology. Given that these measures reflect different genetic mechanisms, clustering based on each individual measure may be advantageous to reveal subgroups that share differences in these mechanisms.

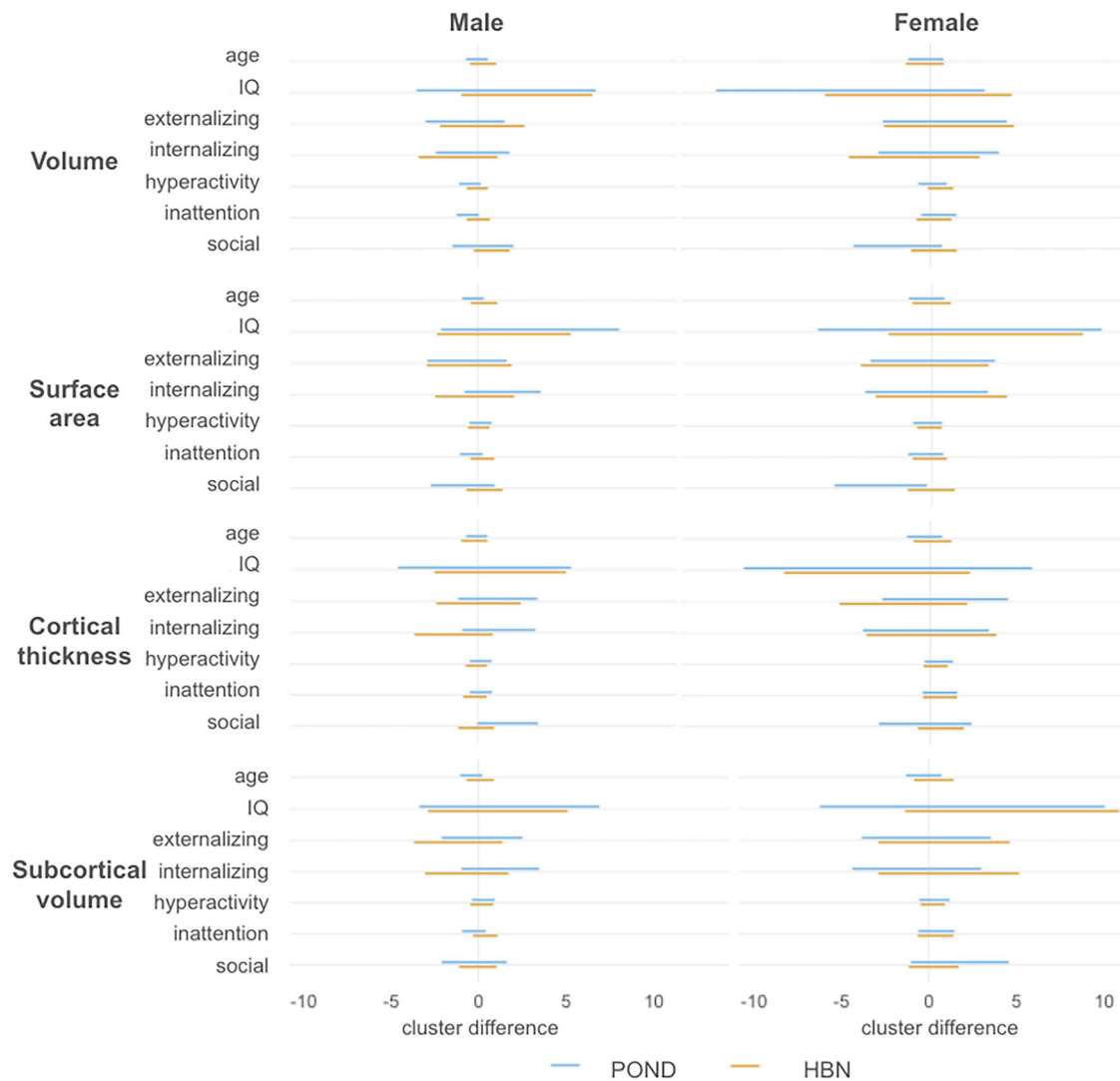
#### Replicability and sex differences

Given the known sex-differences in the neurobiology and phenotypic expression of neurodevelopmental conditions [53], we disaggregated our results by sex. There was high replicability between male and female datasets in the principal component

decomposition of the brain measures, overall clustering structure, lack of alignment with the diagnostic labels, and brain signatures of clusters. In terms of brain-phenotype association, replicability was found in HBN, but not POND. It is important to note that overall, we observed higher variability in the female dataset. This may suggest larger variability in neurobiological characteristics or may be the result of our smaller sample size for the female subsets.

#### STRENGTHS AND LIMITATIONS

This study has several strengths, including our large sample sizes across both datasets. At the same time, there was lower representation of females, matching the expected prevalence in autism and ADHD. This may have limited our ability to detect female-specific patterns. Additionally, our phenotypic measures were limited by what was available in both POND and HBN sets. It may be possible that brain-behaviour associations can be found in other measures of function or cognition (e.g., response inhibition, memory, affect recognition).



**Fig. 3** 95% confidence interval of the mean for the cluster difference reported for phenotypic measures for POND and HBN, disaggregated by sex.

## CONCLUSIONS

To our knowledge, this is the first study of clustering replicability in structural brain measures across neurodevelopmental conditions. We found evidence of replicability of the clustering structure across two independent datasets; however, when examining replicability across brain measures, only replicability across cortical thickness and subcortical volume, and surface area and cortical volume were strongly supported by our results.

## DATA AVAILABILITY

Participants were drawn from the Province of Ontario Neurodevelopmental Disorders (POND) network (exported April 2021; now available via a controlled data release through Ontario Brain Institute's Brain-CODE: <https://www.braincode.ca/>) and the Healthy Brain Network (exported November 2020: [http://fcon\\_1000.projects.nitrc.org/indi/cmi\\_healthy\\_brain\\_network/](http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/)) datasets.

## CODE AVAILABILITY

The codes used in this study are available at <https://github.com/KamranNiroomand/ClusteringAnalysisStudy>.

## REFERENCES

- Christensen DL, Bilder DA, Zahorodny W, Pettygrove S, Durkin MS, Fitzgerald RT, et al. Prevalence and characteristics of autism spectrum disorder among 4-year-old children in the autism and developmental disabilities monitoring network. *J Dev Behav Pediatr*. 2016;37:1–8. <https://doi.org/10.1097/DBP.0000000000000235>.
- Lever AG, Geurts HM. Psychiatric co-occurring symptoms and disorders in young, middle-aged, and older adults with autism spectrum disorder. *J Autism Dev Disord*. 2016;46:1916–30. <https://doi.org/10.1007/s10803-016-2722-8>.
- Thapar A, Cooper M, Rutter M. Neurodevelopmental disorders. *Lancet Psychiatry*. 2017;4:339–46. [https://doi.org/10.1016/S2215-0366\(16\)30376-5](https://doi.org/10.1016/S2215-0366(16)30376-5).
- Boedhoe PSW, van Rooij D, Hoogman M, Twisk JWR, Schmaal L, Abe Y, et al. Subcortical brain volume, regional cortical thickness, and cortical surface area across disorders: findings from the ENIGMA ADHD, ASD, and OCD working groups. *Am J Psychiatry*. 2020;177:834–43. <https://doi.org/10.1176/appi.ajp.2020.19030331>.
- Cao M, Shu N, Cao Q, Wang Y, He Y. Imaging functional and structural brain connectomics in attention-deficit/hyperactivity disorder. *Mol Neurobiol*. 2014;50:1111–23. <https://doi.org/10.1007/s12035-014-8685-x>.
- Kushki A, Anagnostou E, Hammill C, Duez P, Brian J, Iaconi A, et al. Examining overlap and homogeneity in ASD, ADHD, and OCD: a data-driven, diagnosis-agnostic approach. *Transl Psychiatry*. 2019;9:318. <https://doi.org/10.1038/s41398-019-0631-2>.
- Ellegood J, Anagnostou E, Babineau BA, Crawley JN, Lin L, Genestine M, et al. Clustering autism: using neuroanatomical differences in 26 mouse models to gain insight into the heterogeneity. *Mol Psychiatry*. 2015;20:118–25. <https://doi.org/10.1038/mp.2014.98>.

8. Zheng S, Hume KA, Able H, Bishop SL, Boyd BA. Exploring developmental and behavioral heterogeneity among preschoolers with ASD: a cluster analysis on principal components. *Autism Res.* 2020;13:796–809. <https://doi.org/10.1002/aur.2263>.
9. Vandewouw MM, Brian J, Crosbie J, Schachar RJ, Iaboni A, Georgiades S. et al. Identifying replicable subgroups in neurodevelopmental conditions using resting-state functional magnetic resonance imaging data. *JAMA Netw Open.* 2023;6:e232066.
10. Choi EJ, Vandewouw MM, Taylor MJ, Arnold PD, Brian J, Crosbie J, et al. Beyond diagnosis: Cross-diagnostic features in canonical resting-state networks in children with neurodevelopmental disorders. *NeuroImage Clin.* 2020;28:102476. <https://doi.org/10.1016/j.nicl.2020.102476>.
11. Zabihi M, Floris DL, Kia SM, Wolfers T, Tillmann J, Arenas AL, et al. Fractionating autism based on neuroanatomical normative modeling. *Transl Psychiatry.* 2020;10:1–10. <https://doi.org/10.1038/s41398-020-01057-0>.
12. Jacobs GR, Voineskos AN, Hawco C, Stefanik L, Forde NJ, Dickie EW, et al. Integration of brain and behavior measures for identification of data-driven groups cutting across children with ASD, ADHD, or OCD. *Neuropsychopharmacology.* 2021;46:643–53.
13. Chien H, Gau SS, Hsu Y, Chen YJ, Lo YC, Shih YC, et al. Altered cortical thickness and tract integrity of the mirror neuron system and associated social communication in autism spectrum disorder. *Autism Res.* 2015;8:694–708.
14. Fortin JP, Cullen N, Sheline YI, Taylor YD, Aselcioglu Y, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage.* 2018;167:104–20. <https://doi.org/10.1016/j.neuroimage.2017.11.024>.
15. Loth E, Charman T, Mason L, Tillmann J, Jones EJH, Wooldridge C, et al. The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Mol Autism.* 2017;8:24. <https://doi.org/10.1186/s13229-017-0146-8>.
16. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci.* 2017;18:115–26. <https://doi.org/10.1038/nrn.2016.167>.
17. Pipitone J, Park MTM, Winterburn J, Lett TA, Lurch JP, Pruessner JC, et al. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *NeuroImage.* 2014;101:494–512. <https://doi.org/10.1016/j.neuroimage.2014.04.054>.
18. Kushki A, Cardy RE, Panahandeh S, Malihi M, Hammill C, Brian J, et al. Cross-diagnosis structural correlates of autistic-like social communication differences. *Cereb Cortex.* 2021;31:5067–76. <https://doi.org/10.1093/cercor/bhab142>.
19. Committee on Reproducibility and Replicability in Science, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, et al. *Reproducibility and Replicability in Science.* National Academies Press; 2019:25303. <https://doi.org/10.17226/25303>.
20. Becker MM, Wagner MB, Bosa CA, Schmidt C, Longo D, Papaleo C, et al. Translation and validation of Autism Diagnostic Interview-Revised (ADI-R) for autism diagnosis in Brazil. *Arq Neuropsiquiatr.* 2012;70:185–90. <https://doi.org/10.1590/S0004-282X2012000300006>.
21. Scahill L, Riddle MA, McSWIGGIN-HARDIN M, Ort SI, King RA, Goodman WK, et al. Children's Yale-Brown obsessive compulsive scale: reliability and validity. *J Am Acad Child Adolesc Psychiatry.* 1997;36:844–52. <https://doi.org/10.1097/00004583-199706000-00023>.
22. Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, et al. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): initial reliability and validity data. *J Am Acad Child Adolesc Psychiatry.* 1997;36:980–8. <https://doi.org/10.1097/00004583-199707000-00021>.
23. Guha M. Diagnostic and statistical manual of mental disorders: DSM-5. *Ref Rev.* 2014;28:36–37.
24. Rutter M, Bailey A, Lord C. *SCQ. Soc Commun Quest Torrance CA West Psychol Serv.* Published online 2003.
25. Barkley RA. Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychol Bull.* 1997;121:65–94. <https://doi.org/10.1037/0033-2909.121.1.65>.
26. Wechsler D. Wechsler Abbreviated Scale of Intelligence. Published online November 12, 2012. <https://doi.org/10.1037/t15170-000>.
27. Powel J. Wechsler memory scale-revised David A. Wechsler. New York: The Psychological Corporation. Harcourt Brace Jovanovich, Inc, 1987. 150 pp. *Arch Clin Neuropsychol.* 1988;3:397–403. [https://doi.org/10.1016/0887-6177\(88\)90053-4](https://doi.org/10.1016/0887-6177(88)90053-4).
28. Achenbach TM, Edelbrock C. *Child Behavior Checklist.* University Associates in Psychiatry; 2001.
29. Nelson EC, Hanna GL, Hudziak JJ, Botteron KN, Heath AC, Todd RD. Obsessive-compulsive scale of the child behavior checklist: specificity, sensitivity, and predictive power. *Pediatrics.* 2001;108:E14. <https://doi.org/10.1542/PEDS.108.1.E14>.
30. Lepage C, Wagstyl K, Jung B, Seidlitz J, Sponheim C, Ungerleider M, et al. CIVET-Macaque: an automated pipeline for MRI-based cortical surface generation and cortical thickness in macaques. *NeuroImage.* 2021;227:117622. <https://doi.org/10.1016/j.neuroimage.2020.117622>.
31. Park MTM, Pipitone J, Baer LH, Winterburn JL, Shah Y, Chavez S, et al. Derivation of high-resolution MRI atlases of the human cerebellum at 3T and segmentation using multiple automatically generated templates. *NeuroImage.* 2014;95:217–31. <https://doi.org/10.1016/j.neuroimage.2014.03.037>.
32. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging.* 1998;17:87–97. <https://doi.org/10.1109/42.668698>.
33. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* 1987;20:53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
34. Lyttelton O, Boucher M, Robbins S, Evans A. An unbiased iterative group registration template for cortical surface analysis. *NeuroImage.* 2007;34:1535–44. <https://doi.org/10.1016/j.neuroimage.2006.10.041>.
35. Kim JS, Singh V, Lee JK, Lerch J, Ad-Dab'bagh Y, MacDonald D, et al. Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *NeuroImage.* 2005;27:210–21. <https://doi.org/10.1016/j.neuroimage.2005.03.036>.
36. Cauvet É, van't Westeinde A, Toro R, Kuja-Halkola R, Neufeld J, Mevel K, et al. Sex differences along the autism continuum: a twin study of brain structure. *Cereb Cortex.* 2019;29:1342–50. <https://doi.org/10.1093/cercor/bhy303>.
37. Greven CU, Richards JS, Buitelaar JK. *Sex Differences in ADHD.* Vol 1. Oxford University Press; 2018. <https://doi.org/10.1093/med/9780198739258.003.0016>.
38. Lawrence KE, Hernandez LM, Bowman HC, Padgaonkar NT, Fuster E, Jack A, et al. Sex Differences in functional connectivity of the salience, default mode, and central executive networks in youth with ASD. *Cereb Cortex.* 2020;30:5107–20. <https://doi.org/10.1093/cercor/bhaa105>.
39. Schmithorst VJ, Holland SK. Sex differences in the development of neuroanatomical functional connectivity underlying intelligence found using Bayesian connectivity analysis. *NeuroImage.* 2007;35:406–19. <https://doi.org/10.1016/j.neuroimage.2006.11.046>.
40. Tipping ME, Bishop CM. Probabilistic Principal Component Analysis. *J R Stat Soc Ser B Stat Methodol.* 1999;61:611–22. <https://doi.org/10.1111/1467-9868.00196>.
41. Polychronopoulou A, Zhou F, Obradovic Z. Cosine similarity for multiplex network summarization. *In ;* 2021:56–63.
42. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.*; 1988.
43. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Ser B Stat Methodol.* 2001;63:411–23.
44. Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat - Theory Methods.* 1974;3:1–27. <https://doi.org/10.1080/03610927408827101>.
45. LaPlante RA, Douw L, Tang W, Stufflebeam SM. The connectome visualization utility: Software for visualization of human brain networks. *PLoS One.* 2014;9:e113838.
46. Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage.* 2010;52:1059–69.
47. Steinley D. Properties of the hubert-arable adjusted rand index. *Psychol Methods.* 2004;9:386–96. <https://doi.org/10.1037/1082-989X.9.3.386>.
48. Breiman L. Random Forests. *Mach Learn.* 2001;45:5–32. <https://doi.org/10.1023/A:1010933404324>.
49. Sadat-Nejad Y, Vandewouw MM, Cardy R, Lerch JP, Taylor MJ, Iaboni A, et al. Investigating heterogeneity across autism, ADHD, and typical development using measures of cortical thickness, surface area, cortical/subcortical volume, and structural covariance. *Front Child Adolesc Psychiatry.* 2023;2. Accessed February 8, 2024. <https://www.frontiersin.org/articles/10.3389/frcha.2023.1171337>.
50. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoun AS, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature.* 2022;603:654–60. <https://doi.org/10.1038/s41586-022-04492-9>.
51. Panizzon MS, Fennema-Notestine C, Eyler LT, Jernigan TL, Prom-Wormley E, Neale M, et al. Distinct genetic influences on cortical surface area and cortical thickness. *Cereb Cortex.* 2009;19:2728–35. <https://doi.org/10.1093/cercor/bhp026>.
52. Im K, Lee JM, Lyttelton O, Kim SH, Evans AC, Kim SI. Brain size and cortical structure in the adult human brain. *Cereb Cortex.* 2008;18:2181–91. <https://doi.org/10.1093/cercor/bhm244>.
53. Bölte S, Neufeld J, Marschik PB, Williams ZJ, Gallagher L, Lai MC. Sex and gender in neurodevelopmental conditions. *Nat Rev Neurol.* 2023;19:136–59. <https://doi.org/10.1038/s41582-023-00774-6>.

## AUTHOR CONTRIBUTIONS

YSN conceptualized and designed the study, analysed study data, drafted the original manuscript, interpreted the results, and revised the manuscript. MM generated manuscript figures, and critically reviewed and revised the manuscript. EA contributed to the concept and design of the study, acquisition of study data, interpretation of results, and the critical review and revision of the manuscript. JL, MJT, AI, JB, EK, MA, JC, RJS, SG, RN, JJ, PDA, and JPL contributed to the acquisition of data and the critical review and revision of the manuscript. BS and CH contributed to processing of the neuroimaging



data and generation of figures. AK conceptualized and designed the study and analytic plan, interpreted the results, and drafted the manuscript. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

### COMPETING INTERESTS

Dr Nicolson reported receiving grants from Brain Canada, Hoffman La Roche, Otsuka Pharmaceuticals, and Maplight Therapeutics outside the submitted work. Dr Anagnostou reported receiving grants from Roche and Anavex; receiving nonfinancial support from AMO Pharma and CRA-Simons Foundation; and receiving personal fees from Roche, Impel, Ono, and Quadrant outside the submitted work; in addition, Dr Anagnostou had a patent for Anxiety Meter issued 14/755/084 (United States) and a patent for Anxiety Meter pending 2,895,954 (Canada) as well as receiving royalties from APPI and Springer. Dr Kushki reported receiving grants from National Science and Engineering Research Council during the conduct of the study; in addition, Dr Kushki had a patent for Anxiety Meter with royalties paid from Awake Labs. No other disclosures were reported. Dr Kushki is the inventor of a software called holly (formerly "Anxiety Meter"). She is involved in commercialising the holly (patents US 9,844,332 B2) and will financially benefit from its sales. Dr Kushki served on the board of advisors for Shaftesbury, a media company developing virtual reality products for autistic children, from February 2020 to February 2021, and was compensated financially for this role. She has also received consulting fees from DNASTack and donations of hardware for her research program from Samsung Canada.

### ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41398-025-03540-y>.

**Correspondence** and requests for materials should be addressed to Azadeh Kushki.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025