Neural Computing and Applications (2021) 33:3299–3310
https://doi.org/10.1007/s00521-020-05193-y

ORIGINAL ARTICLE

Performance of machine learning classification models of autism using resting-state fMRI is contingent on sample heterogeneity

Maya A. Reiter[1,3] · Afrooz Jahedi[1,2] · A. R. Jac Fredo[1] · Inna Fishman[1] · Barbara Bailey[4] · Ralph-Axel Müller[1,3]

Received: 15 June 2020 / Accepted: 11 July 2020 / Published online: 24 July 2020

Abstract

Autism spectrum disorders (ASDs) are heterogeneous neurodevelopmental conditions. In fMRI studies, including most machine learning studies seeking to distinguish ASD from typical developing (TD) samples, cohorts differing in gender and symptom severity composition are often treated statistically as one 'ASD group.' Using resting-state functional connectivity (FC) data, we implemented random forest to build diagnostic classifiers in four ASD samples including a total of 656 participants ($N_{\mathrm{ASD}} = 306$, $N_{\mathrm{TD}} = 350$, ages 6–18). Groups were manipulated to titrate heterogeneity of gender and symptom severity and partially overlapped. Each sample differed on inclusionary criteria: (1) all genders, unrestricted severity range; (2) only male participants, unrestricted severity; (3) all genders, higher severity only; and (4) only male participants, higher severity. Each set consisted of 200 participants per group (ASD, TD; matched on age and head motion): 160 for training and 40 for validation. FMRI time series from 237 regions of interest (ROIs) were Pearson correlated in a 237×237 FC matrix, and classifiers were built using random forest in training samples. Classification accuracies in validation samples were 62.5%, 65%, 70%, and 73.75%, respectively, for samples 1–4. Connectivity within cingulo-opercular task control (COTC) network, and between COTC ROIs and default mode and dorsal attention network contributed overall most informative features, but features differed across sets. Findings suggest that diagnostic classifiers vary depending on ASD sample composition. Specifically, greater homogeneity of samples regarding gender and symptom severity enhances classifier performance. However, given the true heterogeneity of ASDs, performance metrics alone may not adequately reflect classifier utility.

Keywords Autism spectrum disorder · Autism diagnostic observation schedule · Conditional random forest · Functional connectivity · fMRI · Symptom severity · Machine learning · Heterogeneity

1 Introduction

Autism spectrum disorders (ASDs) encompass neurodevelopmental conditions characterized by sociocommunicative impairments and presence of restricted and

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00521-020-05193-y) contains supplementary material, which is available to authorized users.

✉ Ralph-Axel Müller
rmueller@sdsu.edu

[1] Brain Development Imaging Lab (BDIL), Psychology, San Diego State University (SDSU), 6363 Alvarado Ct. Suite 200, San Diego, CA 92120, USA

[2] Computational Science, San Diego State University/Claremont Graduate University's Joint Doctoral Program, San Diego, CA, USA

[3] Joint Doctoral Program in Clinical Psychology, San Diego State University/UC San Diego, San Diego, CA, USA

[4] Department of Mathematics and Statistics, San Diego State University, San Diego, CA, USA

repetitive patterns of behavior [1]. Although ASDs occur on a spectrum from mild to severe, most individuals require lifelong support in some form (familial, community, taxpayer, health systems, etc.), making research in this population a high public health priority [2]. Despite a large number of neuroimaging studies, no distinct neural signature for ASDs has been established, likely because no single etiology could account for all variants of the disorder [3]. Moreover, many additional factors including age, comorbidities, and treatment history affect neural development; thus, neural signatures for 'classifying' ASDs at the individual level cannot be easily identified. Although findings of atypical brain functioning in ASDs have not fully converged, most studies using functional magnetic resonance imaging (fMRI) have found that children and adults with ASDs show atypical functional connectivity (FC) between many brain regions, both within and between neural networks [4]. FMRI detects low-frequency fluctuations in the blood oxygen level-dependent (BOLD) signal during task performance or in a resting state. Correlations between BOLD time series in different brain regions are considered to reflect FC [5].

Machine learning (ML) classification methods promise to advance understanding of how brain functioning is affected in ASDs at the group level, since they are data-driven approaches that do not require specific hypotheses and can tolerate the high dimensionality of fMRI datasets. Indeed, FC variables have been used in previous ML studies for diagnostic classification of ASD [6–8]. A subsection of these studies (e.g., [9–11]) implemented random forest (RF) [12], a nonparametric ensemble ML method that constructs numerous decision trees (classification and regression trees) and makes final predictions (e.g., ASD vs. non-ASD) based on the class majority votes or mean prediction from all trees. FC variables can be used to build classifiers with RF. Such data-driven approaches can identify atypical brain patterns in ASDs even when specific hypotheses may be unwarranted (given inconsistent evidence in the literature, e.g., [13].

Previous ML studies of ASDs have mostly included highly heterogeneous ASD samples with respect to the range of symptom severities, gender composition, medication status, and other variables. It is currently unknown how sample heterogeneity affects the performance of ML classifiers of ASDs. This restricts the inferences that can be drawn from previous research. Heterogeneity is a critical factor that has hindered progress in understanding atypical brain organization in ASDs in general [14]. For example, although about 20% of individuals with ASDs are female, the impact of gender on atypical brain connectivity in ASDs is not fully understood [15]. The limited literature comparing FC in males and females with ASDs suggests sex differences [15] that are network dependent [16, 17], presumably due to divergent neurodevelopmental trajectories [18]. Secondly, studies often include ASD participants with mild symptom severity, as these may comply more easily with constraints and instruction during rs-fMRI scans. However, since atypical connectivity may be more distinct in ASD individuals with greater symptom severity, inclusion of mild cases may impede diagnostic classification.

Our study aimed to examine the impact of ASD sample heterogeneity on ML diagnostic classification, with respect to both accuracy and selection of informative features. Many variables contribute to heterogeneity in ASDs (e.g., treatment history, language abilities, psychiatric comorbidities). However, given limitations of publicly available large samples, we focused on symptom severity and self/caregiver-reported gender. This focus allowed us to operationalize two factors of heterogeneity that likely affect functional connectivity. We gathered four samples of children and adolescents between the ages of six and eighteen, with varying levels of heterogeneity in gender composition and autism symptom severity, and used resting-state fMRI data to construct diagnostic classifiers for each. We hypothesized that classification accuracy would increase with greater sample homogeneity and that most important FC features would differ across samples.

2 Methods

2.1 Datasets and participants

Children and adolescents between the ages of six and eighteen with ASDs and typically developing (TD) same-aged peers were selected from an in-house sample collected at San Diego State University (SDSU) and from publicly available multisite databases, the autism brain imaging data exchange (ABIDE) I and II [19, 20]. The scanning sites from ABIDE included in the current study were SDSU (in-house, acquired in the period between 2012 and 2019), the Kennedy Krieger Institute (KKI), New York University (NYU), Georgetown University (GU), University of California at Los Angeles (UCLA), University of Utah School of Medicine (USM), and Oregon Health and Science University (OHSU). Supplementary Figures 1–4 illustrate the number of participants from each scanning site included in each sample. Although the in-house dataset is partially available in the ABIDE database, additional scans (unavailable in ABIDE) were included. For in-house data, ASD diagnostic status was established based on the DSM-5 criteria [1], supported by autism diagnostic observation schedule (ADOS) [21] or ADOS-2 [22] administered by research-reliable clinicians, the autism diagnostic interview-revised [23] administered to caregivers, and expert clinical judgment. Diagnostic status for ASD participants in the current study was adopted directly from the ABIDE database (see Supplementary Methods for details). Inclusion in the TD group required absence of ASD or any developmental disorder apart from ADHD. Other specific exclusion criteria used for selecting TD participants (e.g., absence of neurological or psychiatric history or absence of family history of ASDs) varied across ABIDE sites. Due to known effects of eye status on FC [24], only resting-state fMRI datasets acquired with participants' eyes open were included. As even minute amounts of head motion are known to impact FC estimates [25], strict head motion criteria were adopted. Only participants with $\geq$ 80% of volumes retained after censoring volumes with framewise displacement $\geq$ 0.5 mm and root-mean-square displacement (RMSD, a summary measure of head motion throughout the scan) < 0.2 mm were included. Further, only ABIDE sites contributing $\geq$ 10 participants meeting these head motion inclusion criteria were included; 306 participants with ASD and 350 TD participants were included. The fMRI and clinical data shared in the ABIDE database were de-identified from protected health information. Collection of all in-house data was approved by the Institutional Review Boards at San Diego State University and the University of California, San Diego, and all participants provided written consent to participate and also to have their data included in the ABIDE databases. For ABIDE data, prior to data contribution, sites were required to confirm that their local Institutional Review Board or ethics committee had approved both the initial data collection and the sharing of fully de-identified datasets.

Four classification samples (each including 200 ASD and 200 TD participants) were created by constricting ASD sample heterogeneity in gender and ADOS or ADOS-2 total scores (depending on the edition of which was ADOS administered). The four samples were as follows: sample 1 ('full heterogeneity'): participants with a full range (unrestricted) of ADOS Total scores, gender unrestricted; sample 2 ('reduced gender heterogeneity'): males only, all ADOS Total scores included; sample 3 ('reduced ASD-symptom heterogeneity'): gender unrestricted with ADOS Total scores $\geq$ 10; and sample 4 ('low heterogeneity'): males only with ADOS scores $\geq$ 10 (see Supplemental Methods for details on ADOS Total scores).

Given the overall small number of female participants in ABIDE, corresponding to the generally lower prevalence of ASDs in females [26], female-only samples could not be included. Tables 1, 2, 3, and 4 present the demographic characteristics of training and validation datasets for samples 1–4. In each sample, ASD and TD participants were matched, at the group level, on head motion, handedness, and age. We also matched groups on gender (where appropriate) and performance intelligence quotient (PIQ) to the extent possible given sample limitations. Matching was carried out using an algorithm developed by coauthor A.J. [27], which determines the optimal overall multivariate matching solution for two groups of participants by using proximity or distance-based metrics across multiple variables (here: age, motion, PIQ, and in samples 1 and 3, gender). Where PIQ was not available from ABIDE data (sample 1: $n = 58$, sample 2: $n = 58$, sample 3: $n = 67$, sample 4: $n = 66$), full-scale intelligent quotient (FSIQ) was used for matching purposes. In a proportion of 80:20, datasets were selected randomly from each site for the inclusion in the training model and validation models (160 ASD participants for training model and 40 ASD participants for the validation model). ASD participants included in the training and validation models were separately matched to same-sized TD groups for training and validation using the matching algorithm described above. Sample overlap for the four comparisons is depicted using Venn diagrams in Supplementary Figure 4.

2.2 Data acquisition and preprocessing

For each participant, anatomical MRI and resting-state fMRI scans were collected. Acquisition parameters for each included scanning site are detailed in Supplementary Tables 1 and 2. Resting-state fMRI images were preprocessed using the Analysis of Functional NeuroImages, AFNI, [28] (afni.nimh.nih.gov) and FSL 5.0 [29] (fmrib.ox.ac.uk/fsl) software packages. Volumes at the beginning of each scan were discarded for T1 equilibration as needed (see Supplementary Table 3). In-house resting-state fMRI data were field-map and slice-time corrected. All data were motion-corrected and resampled to MNI 3 mm isotropic standard space using normalization tools from FSL: FLIRT (functional to anatomical, six degrees of freedom) and FNIRT (anatomical to standard, using sinc interpolation). Images were spatially smoothed to a global full width at half maximum of 6 mm and temporally smoothed using a $0.008 < f < 0.08$ Hz band-pass filter. Subject-level regression of eight nuisance variables and their first derivatives (six rigid-body motion parameters estimates from motion correction, mean white-matter signal, and ventricular cerebrospinal fluid obtained via FSL's FAST) was performed for denoising. Global signal regression was additionally performed, as recommended for multisite data [30]. All seventeen nuisance regressors were band-pass filtered by the same second-order Butterworth filter ($0.008 < f < 0.08$ Hz) as the functional time series [31]. Single volumes with framewise displacement > 0.5 mm and two immediately following volumes were censored, and time series segments with fewer than 10 contiguous volumes after censoring were also removed. Preprocessed images were visually inspected for artifacts and suitable brain coverage.

Table 1 Participant demographics for sample 1 ('full heterogeneity': all gender participants with unrestricted ADOS Total scores)

	ASD, mean ± SD (range)	TD, mean ± SD (range)	*p* value (*t* or X^2 test)
Training model sample N (handedness)	160 (R-107, L-12, M-11, ND-30)	160 (R-118, L-10, M-6, ND-26)	–
Age (years)	12.15 ± 3.04 (6.41–18)	12.02 ± 2.77 (6.36–18.55)	0.68
In-scanner motion (mm)	0.095 ± 0.04 (0.021–0.197)	0.09 ± 0.039 (0.02–0.188)	0.31
PIQ/FSIQ	105.8 ± 17.88 (53–149)	111.25 ± 13.11 (67–138)	0.002
Females	*n* = 24	*n* = 38	0.047
Validation sample N (handedness)	40 (R-27, L-1, M-1, ND-11)	40 (R-26, L-1, M-1, ND-12)	–
Age (years)	12.02 ± 2.96 (7.15–17.53)	12.78 ± 2.52 (8.74–17.1)	0.22
Motion (mm)	0.09 ± 0.043 (0.036–0.172)	0.083 ± 0.034 (0.03–0.177)	0.44
PIQ/FSIQ	106.1 ± 17.4 (81–145)	109.2 ± 13.93 (62*–131)	0.38
Females	*n* = 5	*n* = 7	0.53

SD standard deviation, *N* number of participants, *R* right, *L* left, *M* mixed, *ND* no data

*The TD participant with PIQ 62 had a FSIQ in the low-average range (FSIQ = 79. Verbal IQ = 101)

Table 2 Participant demographics for sample 2 ('reduced gender heterogeneity': males only with unrestricted ADOS Total scores)

	ASD, mean ± SD (range)	TD, mean ± SD (range)	*p* value (2 sample *t* test)
Training model sample N (handedness)	160 (R-108, L-11, M-8, ND-33)	160 (R-121, L-9, M-5, ND-25)	–
Age (years)	12.16 ± 2.76 (6.41–17.94)	12.03 ± 2.9 (6.36–18.8)	0.7
Motion (mm)	0.094 ± 0.041 (0.021–0.191)	0.09 ± 0.037 (0.026–0.191)	0.32
PIQ/FSIQ	106.8 ± 16.67 (69–146)	111.9 ± 12.78 (83–147)	0.002
Validation sample N (handedness)	40 (R-24, L-1, M-1, ND-14)	40 (R-23, L-2, M-3, ND-12)	–
Age (years)	12.06 ± 2.67 (7.15–17.5)	12.82 ± 2.57 (8.06–18.21)	0.19
Motion (mm)	0.088 ± 0.41 (0.035–0.172)	0.08 ± 0.045 (0.027–0.177)	0.43
PIQ/FSIQ	105.5 ± 17.19 (77–145)	111.8 ± 16.17 (62*–147)	0.095

SD standard deviation, *N* number of participants, *R* right, *L* left, *M* mixed, *ND* no data

*The TD participant with PIQ 62 had a FSIQ in the low-average range (FSIQ = 79. Verbal IQ = 101)

Table 3 Participant demographics for sample 3 ('reduced ASD-related heterogeneity': all genders and ADOS Total scores ≥ 10)

	ASD, mean ± SD (range)	TD, mean ± SD (range)	*p* value (*t or* chi^2 test)
Training model sample N (handedness)	160 (R-111, L-12, M-9, ND-28)	160 (R-120, L-11, M-5, ND-24)	–
Age (years)	11.9 ± 2.69 (6.41–18)	11.7 ± 2.62 (6.36–18.55)	0.68
Motion (mm)	0.097 ± 0.04 (0.031–0.197)	0.092 ± 0.041 (0.023–0.191)	0.29
PIQ/FSIQ	105.1 ± 17.16 (53–146)	111.6 ± 13.3 (77–139)	0.0001
Females	*n* = 23	*n* = 33	0.141
Validation sample N (handedness)	40 (R-24, L-1, M-2, ND-13)	40 (R-30, L-1, M-1, ND-8)	–
Age (years)	12.53 ± 2.89 (8.78–18.58)	11.79 ± 2.78 (7.25–17.5)	0.24
Motion (mm)	0.092 ± 0.038 (0.027–0.154)	0.084 ± 0.046 (0.021–0.177)	0.41
PIQ/FSIQ	110.05 ± 16.38 (82–142)	109.3 ± 14.65 (62*–140)	0.82
Females	*n* = 6	*n* = 12	0.108

SD standard deviation, *N* number of participants, *R* right, *L* left, *M* mixed, *ND* no data

*The TD participant with PIQ 62 had a FSIQ in the low-average range (FSIQ = 79. Verbal IQ = 101)

Table 4 Participant demographics for sample 4 ('low heterogeneity': males only with ADOS Total scores ≥ 10)

	ASD, mean ± SD (range)	TD, mean ± SD (range)	p value (2 sample t test)
Training model sample N (handedness)	160 (R-106, L-13, M-7, ND-34)	160 (R-116, L-10, M-3, ND-31)	–
Age (years)	12.25 ± 2.84 (6.41–18.65)	12.13 ± 2.87 (6.66–18.8)	0.71
Motion (mm)	0.096 ± 0.04 (0.017–0.197)	0.091 ± 0.04 (0.027–0.191)	0.31
PIQ/FSIQ	105.45 ± 16.77 (67–146)	110.65 ± 12.92 (76–138)	0.002
Validation sample *N* (handedness)	40 (R-24, L-1, M-3, ND-12)	40 (R-25, L-1, M-1, ND-13)	–
Age (years)	12.93 ± 2.81 (8.06–18.15)	12.11 ± 2.67 (8–18.24)	0.18
Motion (mm)	0.082 ± 0.043 (0.03–0.17)	0.089 ± 0.043 (0.027–0.19)	0.44
PIQ/FSIQ	109.8 ± 16.74 (62–147)	106.9 ± 15.66 (82–137)	0.43

SD standard deviation, *N* number of participants, *R* right, *L* left, *M* MIXED, *ND* no data

2.3 Regions of interest and connectivity matrices

Mean time series were extracted from 237 cortical, subcortical, and cerebellar regions of interest (ROIs) and used to create a 237 × 237 Fisher-transformed Pearson correlation connectivity matrix for each participant. This final set of ROIs included 214 of 333 ROIs from the cortical atlas by Gordon et al. [32] and all 14 subcortical, as well as 9 of 26 cerebellar ROIs from the Harvard–Oxford subcortical [33] and cerebellar [34] atlases (combining several small ROIs). The following procedure was used to determine inclusion of each ROI: First, we identified all voxels of the brain which included BOLD signal in > 95% of participants. Next, we calculated the percentage of these included voxels in each ROI. Only ROIs for which > 95% of voxels contained true BOLD signal in > 95% of participants were included; 119 cortical and 17 cerebellar ROIs did not meet these criteria and were excluded. The connectivity matrices were calculated separately for each participant in the training and validation set, with a size of 320 × 27,966 and 80 × 27,966, respectively.

2.4 Machine learning algorithm

Random forest (RF) is an ensemble learning classifier that builds a set of classification trees using a training dataset and then classifies the validation dataset by the majority vote of the predictors. RF is superior to other ensemble methods, such as bagging and boosting, in terms of accuracy, computational time, robustness against overfitting, and user interface to choose tree size [12]. It has been suggested that conditional random forest (CRF), which mitigates problems tied to correlated predictor variables [35] as present in fMRI, performs better in feature reduction, whereas RF is preferable for classification [11]. Therefore, feature selection from the high-dimensional FC matrix (320 × 27,966) was carried out with CRF [11], and the classification model was built with RF, using packages from R [36].

In our study, using the RF algorithm the training data were obtained from a bootstrap sample of the data where approximately 2/3 of the data was included in the training dataset and the remaining 1/3, known as out-of-bag (OOB) data, was used to test the error of the predictions for each tree. The RF algorithm was initialized with the number of trees grown and the number of variables used for the split at each node. An un-pruned tree from each bootstrap sample was grown such that at each node a number of predictors were randomly selected as a subset of predictor variables, and the best split from among those variables was chosen [12]. The splitting algorithm used in RF is the GINI index criterion. The OOB data were predicted and corresponding error calculated based on the outcome of the predictions of the ensemble of trees, after all the trees were grown [37].

The GINI measure can identify the influential predictor variables in random forest, but the influence of correlated and uncorrelated predictor variables is better identified using the conditional permutation importance [38]. We used CRF dimensional reduction algorithm [11] with 2001 trees, to calculate the conditional permutation importance (CPI). The most informative features based on sorted positive values of CPI were selected. These most informative features were then fed into CRF dimensional reduction method to find the next set of significant features. We repeated this procedure until the feature set was reduced to 2. The optimal feature set for the model was identified based on the lowest OOB error in the training dataset. We evaluated the performance of the CRF model with optimal features in the test set using the RF classifier with a tree number of 20,000. The number of trees for the CRF and RF was chosen based on stability of results in both CRF and RF. The number of variables at each node was optimized as the square root of overall variables. The performance of the model using the test set was evaluated using the metrics such as accuracy, sensitivity, and specificity. The accuracy depicts the ability of the model to differentiate the ASD and TD correctly and is given by

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The sensitivity shows the ability of the model to determine the ASD correctly and calculated by

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The specificity represents the ability of the model to determine the TD correctly and estimated by

$$\text{Accuracy} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

where TP is the true positive (the number of participants correctly identified as ASD), FP is the false positive (the number of participants incorrectly identified as ASD), TN is the true negative (the number of participants correctly identified as TD), and FN is the false negative (the number of participants incorrectly identified as healthy). The informative brain networks were identified from the features that helped to achieve peak classification accuracy.

3 Results

The performance of the classifiers over a range of feature sets, using the training data, is shown in Fig. 1a–d. Feature sets > 1024 features failed to produce higher accuracies and are not shown. The feature set achieving the lowest OOB error rate for the training model was selected, and its performance was tested in the validation dataset; classification results for the validation test are summarized in Table 5. This table indicates the influence of ASD-symptom severity and gender composition on classification accuracy, sensitivity, specificity, and number of features contributing to the best classification performance in the validation samples. Peak classification accuracies in the training models are also included. Reduced sample heterogeneity was overall associated with increased classification accuracy, sensitivity, and specificity. Figure 2a–d shows for each of samples 1–4 the proportion of ROIs included in the classifier that achieved peak classification accuracy from each of the brain networks included in the present study (classified according to atlases described in Sect. 2.3). For each sample, the specific neural networks that contributed the highest proportion of ROIs included in the diagnostic classifier are described in the sections below. Figure 3a–d shows, for each of samples 1–4, the percentage of ROIs participating in most informative connectivities per brain network.

3.1 Sample 1: Full heterogeneity (male and female participants with unrestricted ADOS scores)

OOB error rates of the RF classifier on the training data for the different feature sets are shown in Fig. 1a. A maximum training accuracy of 83.7% (16.2% of OOB error) was achieved with the training model using 72 FC features. For validation sample 1, we obtained a classification test accuracy, sensitivity, and specificity of 62.5%, 60%, and 65%, respectively. Most informative connectivities for the classifier with peak accuracy predominantly involved ROIs in cingulo-opercular task control (COTC), visual, dorsal attention (DA), and somatosensory motor hand (SMH) networks, which together accounted for $> 50\%$ of ROI participations (Fig. 2a).

3.2 Sample 2: Reduced gender heterogeneity (males only with unrestricted ADOS scores)

Figure 1b shows OOB error rate, sensitivity, and specificity for the sample 2 training data. The training model discriminated ASD from TD participants with a peak accuracy of 86.9% (13.1% of OOB error) using 143 FC features. The validation set test accuracy, sensitivity, and specificity of the classifier were all 65%. Prominent among the informative regions were ROIs within COTC, DA, visual, and fronto-parietal task control (FPTC) networks, which together accounted for $> 50\%$ of ROI participations. Figure 3b shows that the highest number of informative classification features (6 features) was found for FC between ROIs of COTC and DA networks.

3.3 Sample 3: Reduced ASD-symptom heterogeneity (all genders, ADOS scores ≥ 10)

The performance of the RF classifier using the training dataset of male and female participants with ADOS Total scores ≥ 10 is shown in Fig. 1c. For this sample, the training model achieved a peak accuracy of 87.2% (12.8% of OOB error) with 324 FC features. In the validation sample, accuracy, sensitivity, and specificity of classification were 70%, 67.5%, and 72.5%, respectively. Many informative FC features involved ROIs within the COTC, DA, visual, ventral attention (VA) and SMH networks, together accounting for 57.8% of ROI participations. Figure 3c shows that connectivities within the COTC (11 informative features) were among the most informative variables for diagnostic classification in this sample.

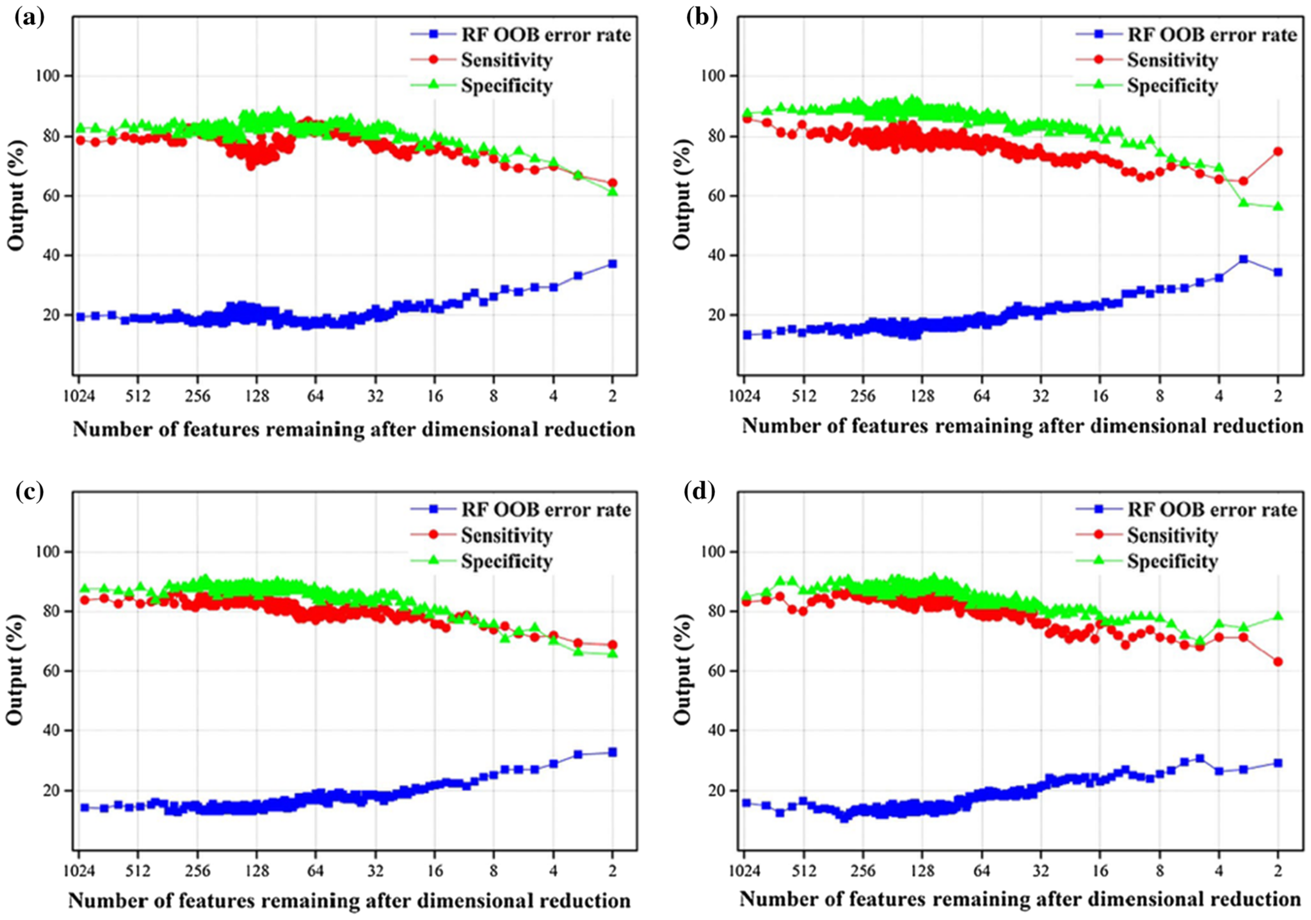

Fig. 1 RF OOB error rate, sensitivity and specificity after each dimension reduction for the training sample. Panels **a**–**d** correspond to samples 1–4, respectively

Table 5 Peak classification accuracy, sensitivity, specificity, number of features and top networks, and connections for samples 1–4

Sample	Acc (train) (%)	Features #	Acc (val) (%)	Sen (val) (%)	Spec (val) (%)	Top four networks (val)	Top connection (val)
1: ADOS total unrestricted, males and females	83.7	72	62.5	60	65	COTC, visual, DA, SMH	DA-SMH
2: ADOS total unrestricted, males only	86.9	143	65	65	65	COTC, DA, visual, FPTC	COTC-DA
3: ADOS total $\geq$ 10, gender unrestricted	87.2	324	70	67.5	72.5	COTC, DA, visual, VA, SMH	Within COTC
4: ADOS total $\geq$ 10, males only	89.4	319	73.7	75	72.5	COTC, visual, SMH, DMN	COTC-DMN and subcortical COTC

Acc accuracy, *Sen* sensitivity, *Spec* specificity, *val* validation set, *train* training set

3.4 Sample 4: Low heterogeneity (males only with ADOS scores ≥ 10)

Training results for sample 4 are shown in Fig. 1d. The performance of the training model improved for the classification of male-only ASD participants with higher symptom severity scores from TD individuals, compared to other samples, with an accuracy of 89.4% (10.6% of OOB error) using 319 features. We also achieved the highest validation test accuracy of 73.7% in this sample. Sensitivity and specificity for this validation model were 75% and 72.5%, respectively. Features including ROIs from

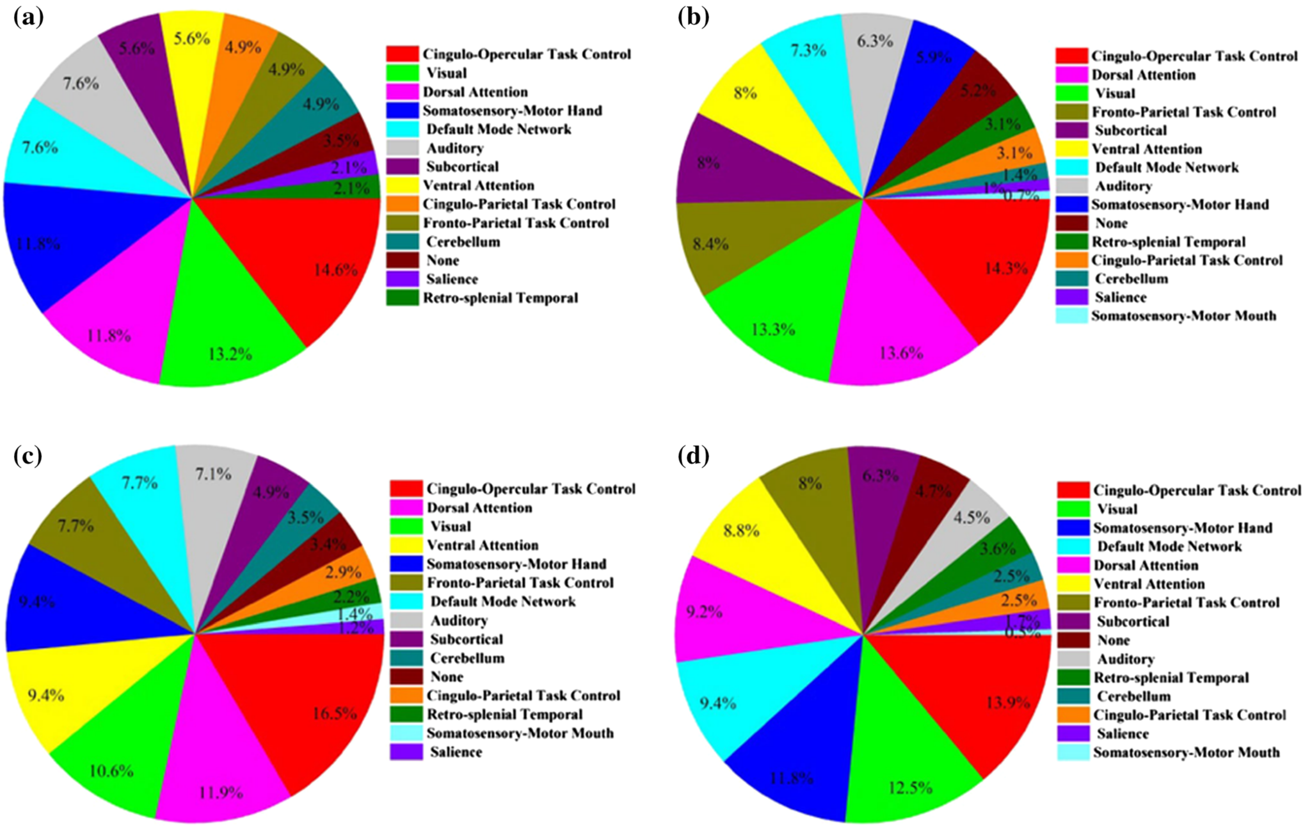

Fig. 2 Pie charts show the brain connections that helped to achieve peak accuracy in each of the four sample sets, separated by network. **a** Full heterogeneity, **b** reduced gender heterogeneity, **c** reduced ASD-symptom heterogeneity, **d** low heterogeneity

COTC, visual, SMH, default mode network (DMN) and DA contributed most to the classifier, accounting for a total of 56.8% ROI participations (Fig. 2d). Connections between ROIs in the COTC and DMN networks and between subcortical and COTC regions were most frequent in the classifier (Fig. 3d).

4 Discussion

4.1 Relationships between ASD sample heterogeneity and diagnostic accuracy

The current study sought to examine how ASD sample heterogeneity with respect to gender and current autism symptom severity impacts performance of ML classifiers built using FC features from resting-state fMRI data. Overall, we found that ML diagnostic classifiers and their accuracy were sensitive to sample heterogeneity. Specifically, we found that FC features maximally distinguishing between ASD and TD groups differed depending on sample composition. Classification accuracy for the validation test in sample 1 (with the greatest heterogeneity in gender composition and ASD-symptom severity) was lowest (62.5%) compared to all other sample sets. With more restricted sample heterogeneity (both in gender composition and in symptom severity), classification accuracy, sensitivity, and specificity increased.

The optimal number of FC features achieving peak accuracy was much higher in the more homogeneous sample 4 (319 features), compared to sample 1 with full heterogeneity (72 features). This was unexpected and is not readily interpretable. One possibility is that the number of FC features that can be used to discriminate between ASD and TD youth reflects commonalities in brain function among children and adolescents in the ASD group that are not shared by TD peers. Thus, an increased number of features could suggest that homogeneity of gender and symptom severity within the ASD sample is associated with greater homogeneity in atypical brain function, corresponding to a greater number of features informing diagnostic classification. Conversely, our results suggest that as sample heterogeneity increases, brain connectivity varies more within the ASD cohort and only a smaller set of shared connectivity features contributes to peak classification accuracy. Our results suggest that observed patterns of FC differences between ASD and TD cohorts are highly sensitive to cohort composition. This is consistent

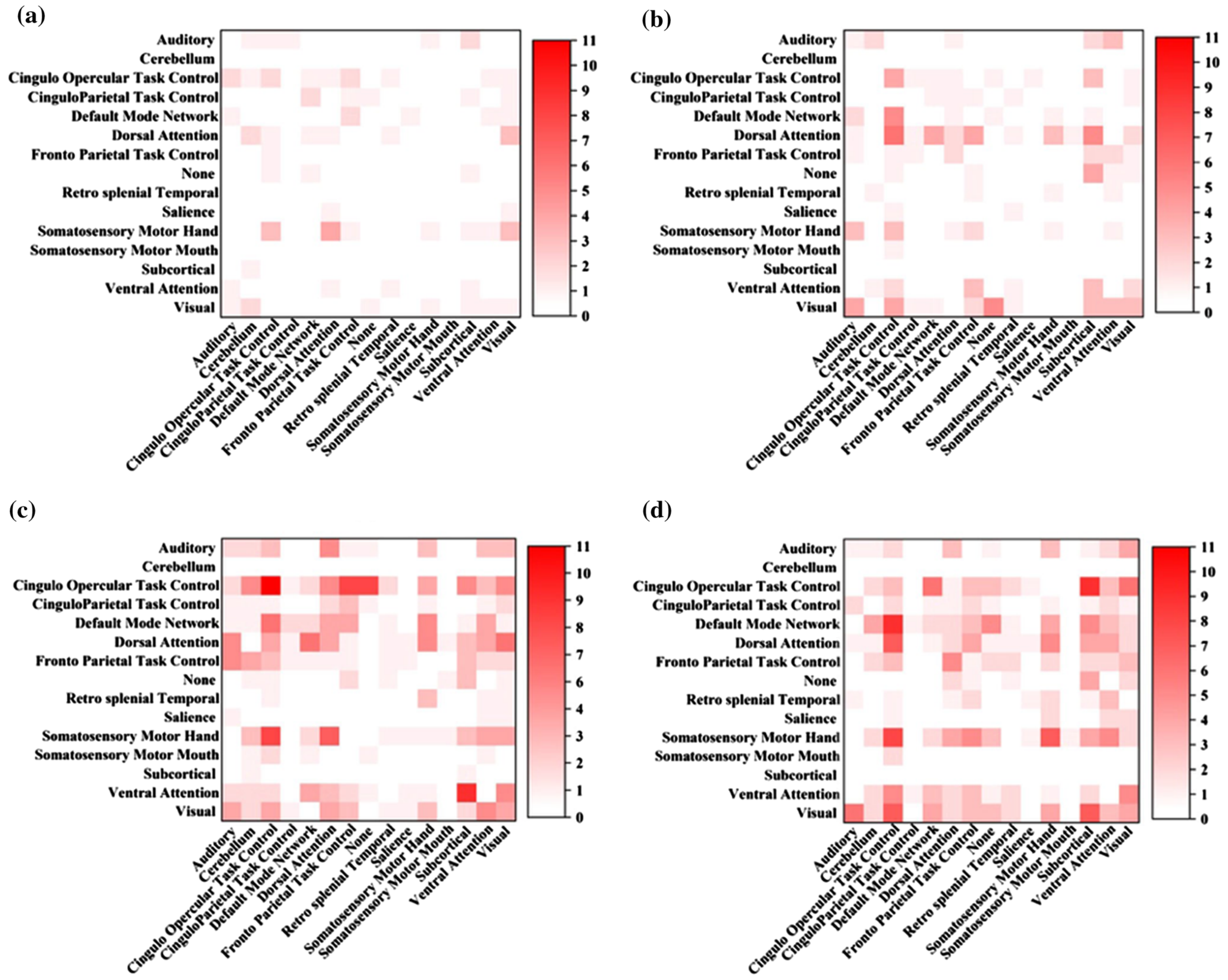

Fig. 3 Heat maps show the number of connectivity features included in the diagnostic classifier between ROIs classified into brain networks. Panels **a–d** correspond to samples 1–4

with recent reports of low replication in resting-state FC across ABIDE sites—and therefore across ASD cohorts [13, 39].

In the present study, we achieved the highest classification accuracy in external validation (73.75%) for sample 4, which included only males with ADOS Total scores ≥ 10. This classifier's accuracy was higher compared to most previous fMRI ML studies of ASD [11, 40, 41] that applied classifiers to completely new external validation data in order to rule out a 'peeking' problem (which occurs when feature selection includes data that are also used for cross-validation) [42]; however, some studies have reported slightly higher classification accuracies [43]. Although achieving higher accuracy in classification is important for understanding how brain function may be different in ASDs, the impact of sample heterogeneity implies that classification accuracy may not be the sole determinant of a classifier's usefulness. A lower classification accuracy obtained in a study including a highly heterogeneous sample that is representative of the diverse autism population at large may be informative in different ways, by pinpointing brain features that tend to be atypical across many (but not all) variants of ASDs. Future research using larger sample sizes that are more heterogeneous, and research aiming to identify subtypes of ASDs from representative samples, including sufficient samples of females and individuals with lower cognitive abilities, is warranted.

We observed that the accuracy of classification decreased as the dimensionality of the feature set was reduced to below 20 (Fig. 1). This suggests that information that is critical for diagnostic classification is lost when models are oversimplified, which may hinder efforts toward discovery of more robust FC biomarkers of autism [44]. In this context, it is remarkable that peak accuracy for most homogeneous sample 4 was achieved for a comparatively high number of over 300 features, which indicates that greater homogeneity at the behavioral and demographic levels may still be associated with heterogeneity at the neural level. Although classifiers with high diagnostic accuracy are important, models containing very large numbers of features may be difficult to interpret. Trade-offs between maximizing prediction accuracy and conceptual interpretation of findings must therefore be considered. In view of the common terminology of 'diagnostic classification,' it is important to remember that ML serves as a research tool for improved understanding of atypical brain functioning in ASDs, whereas applications as a truly diagnostic tool are currently (and possibly in principle) unrealistic for a clinically defined disorder.

4.2 Regions and networks most informative in diagnostic classification

The COTC network contributed the highest number of features to classifiers achieving peak accuracy, across all four samples. This informative role in classification was similar to some recent diagnostic prediction studies of ASD [10, 43]. However, representation of ROIs from other networks varied by sample set. We observed a higher number of informative features involving the COTC and DMN in samples including individuals with ADOS scores indicative of moderate to severe levels of symptoms (samples 3 and 4), compared to samples of participants with a wider range of symptom severity, including individuals with lower ADOS scores (samples 1 and 2). This finding may be related to a significant positive correlation between the ADOS and FC between COTC and DMN observed in a previous study [9]. Relatively high numbers of features from the visual network also contributed to classifiers across all samples. This finding adds to a growing body of research implicating atypical functioning of the visual regions in ASDs [45–48].

Finally, we found that DA and VA network ROIs contributed more features in sample sets 3 and 4, suggesting disruption of connectivity involving attention networks that may help differentiate more severely affected youth with ASDs from TD peers. These findings are also consistent with earlier reports that connectivity in DA and VA may help to differentiate between cohorts of ASD and TD individuals [49].

4.3 Limitations

Low availability of sufficiently large samples of low-motion fMRI data from a multisite source like ABIDE presents unavoidable constraints. Lack of an all-female comparison group was an unavoidable limitation. As further acknowledged in Sect. 1, only two prominent factors of heterogeneity in ASD samples were manipulated in our study. For example, age-related heterogeneity likely impacts fMRI findings of group differences [50]. In the present study, we matched samples on age keeping heterogeneity in age similar across all comparison groups (without being able to eliminate age-related heterogeneity). We used ADOS Total scores as indices of autism symptomatology, as this was the sole relevant measure available across ABIDE sites. ASD symptomatology does not straightforwardly relate to overall cognitive ability or other variables, such as executive functioning. Although we tried to match groups on IQ, the effect of heterogeneity on diverse cognitive and behavioral variables remained beyond the scope of our study, as such collateral data are limited and reported inconsistently across different sites contributing to ABIDE.

5 Conclusions

We examined the impact of sample heterogeneity on diagnostic classification of ASD. Overall, reduced heterogeneity with respect to gender and range of symptom severity was associated with improved performance of RF classifiers. Greater homogeneity of samples impacted both classification accuracy and selection of most informative features. Differences in sample heterogeneity may account for often conflicting findings in the neuroimaging literature on ASDs. Stratification by gender, symptom severity, age, cognitive ability, and other factors of variability may be critical in future efforts to pinpoint atypical brain features of ASDs.

Acknowledgments This study was supported by the National Institutes of Health R01 MH081023 (RAM) and R01 MH101173 (RAM). Authors MAR and AJ were supported by San Diego State University Graduate Fellowships. MAR was also supported by the Autism Speaks Sir Dennis Weatherstone Fellowship program (Grant No. 10609). ARJF was supported by the Science Engineering and Research Board (SERB)/Indo-Us Science and Technology Forum (IUSSTF), Grant 119. We would like to acknowledge here the participants and families who enrolled in the current study, without whom the current project would not have been possible.

Compliance with ethical standards

Conflicts of interest The authors have no known conflicts of interest.

Ethics approval All data collection and sharing were conducted under the approval of the appropriate Institutional Review Boards.

Consent to participate All participants provided written consent to participate and share their de-identified data for the purposes of research.

Consent for publication All authors were involved in the preparation of this manuscript and consent to its publication.

Availability of data and material All of the data included in the current study is publically available through the Autism Brain Imaging Data Exchange (ABIDE) initiative online.

Code availability A combination of publically available neuroimaging processing software, statistical analysis software, and custom code generated by the research team was used in the preparation of this manuscript.

References

1. American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders: Dsm-5tm, 5th ed. American Psychiatric Publishing, Inc
2. Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J (2018) Autism spectrum disorder. Lancet 392(10146):508–520
3. Lenroot RK, Yeung PK (2013) Heterogeneity within autism spectrum disorders: what have we learned from neuroimaging studies? Front Hum Neurosci 7:733
4. Hull JV, Jacokes ZJ, Torgerson CM, Irimia A, Van Horn JD (2017) Resting-state functional connectivity in autism spectrum disorders: a review. Front Psychiatry 7:205
5. Zhang S, Li X, Lv J, Jiang X, Guo L, Liu T (2016) Characterizing and differentiating task-based and resting state fmri signals via two-stage sparse representations. Brain Imaging Behav 10(1):21–32
6. Mastrovito D, Hanson C, Hanson SJ (2018) Differences in atypical resting-state effective connectivity distinguish autism from schizophrenia. Neuroimage Clin 18:367–376
7. Wang CH, Xiao ZY, Wu JH (2019) Functional connectivity-based classification of autism and control using svm-rfecv on rs-fmri data. Phys Med Eur J Med Phys 65:99–105
8. Sen B, Borle NC, Greiner R, Brown MRG (2018) A general prediction model for the detection of adhd and autism using structural and functional mri. PLoS ONE 13(4):e0194856
9. Chen H, Duan XJ, Liu F, Lu FM, Ma XJ, Zhang YX, Uddin LQ, Chen HF (2016) Multivariate classification of autism spectrum disorder using frequency-specific resting-state functional connectivity-a multi-center study. Prog Neuropsychopharmacol Biol Psychiatry 64:1–9
10. Feczko E, Balba NM, Miranda-Dominguez O, Cordova M, Karalunas SL, Irwin L, Demeter DV, Hill AP, Langhorst BH, Painter JG, Van Santen J, Fombonne EJ, Nigg JT, Fair DA (2018) Subtyping cognitive profiles in autism spectrum disorder using a functional random forest algorithm. Neuroimage 172:674–688
11. Jahedi A, Nasamran CA, Faires B, Fan J, Muller RA (2017) Distributed intrinsic functional connectivity patterns predict diagnostic status in large autism cohort. Brain Connect 7(8):515–525
12. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
13. King JB, Prigge MBD, King CK, Morgan J, Weathersby F, Fox JC, Dean DC 3rd, Freeman A, Villaruz JAM, Kane KL, Bigler ED, Alexander AL, Lange N, Zielinski B, Lainhart JE, Anderson JS (2019) Generalizability and reproducibility of functional connectivity in autism. Mol Autism 10:27
14. Lombardo MV, Lai MC, Baron-Cohen S (2019) Big data approaches to decomposing heterogeneity across the autism spectrum. Mol Psychiatry 24(10):1435–1450
15. Alaerts K, Swinnen SP, Wenderoth N (2016) Sex differences in autism: a resting-state fmri investigation of functional brain connectivity in males and females. Soc Cogn Affect Neurosci 11(6):1002–1016
16. Floris DL, Lai MC, Nath T, Milham MP, Di Martino A (2018) Network-specific sex differentiation of intrinsic brain function in males with autism. Mol Autism 9:17
17. Olson LA, Mash LE, Linke A, Fong CH, Mueller RA, Fishman I. Sex-related patterns of intrinsic functional connectivity in children and adolescents with autism spectrum disorders. Autism (in press)
18. Smith REW, Avery JA, Wallace GL, Kenworthy L, Gotts SJ, Martin A (2019) Sex differences in resting-state functional connectivity of the cerebellum in autism spectrum disorder. Front Hum Neurosci 13:104
19. Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, Balsters JH, Baxter L, Beggiato A, Bernaerts S, Blanken LM, Bookheimer SY, Braden BB, Byrge L, Castellanos FX, Dapretto M, Delorme R, Fair DA, Fishman I, Fitzgerald J, Gallagher L, Keehn RJ, Kennedy DP, Lainhart JE, Luna B, Mostofsky SH, Muller RA, Nebel MB, Nigg JT, O'Hearn K, Solomon M, Toro R, Vaidya CJ, Wenderoth N, White T, Craddock RC, Lord C, Leventhal B, Milham MP (2017) Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci Data 4:170010
20. Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, Anderson JS, Assaf M, Bookheimer SY, Dapretto M, Deen B, Delmonte S, Dinstein I, Ertl-Wagner B, Fair DA, Gallagher L, Kennedy DP, Keown CL, Keysers C, Lainhart JE, Lord C, Luna B, Menon V, Minshew NJ, Monk CS, Mueller S, Muller RA, Nebel MB, Nigg JT, O'Hearn K, Pelphrey KA, Peltier SJ, Rudie JD, Sunaert S, Thioux M, Tyszka JM, Uddin LQ, Verhoeven JS, Wenderoth N, Wiggins JL, Mostofsky SH, Milham MP (2014) The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19(6):659–667
21. Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L, Schopler E (1989) Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. J Autism Dev Disord 19(2):185–212
22. Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop SL (2012) Autism diagnostic observation schedule, 2nd edn. Western Psychological Services, Torrance
23. Lord C, Rutter M, Le Couteur A (1994) Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord 24(5):659–685
24. Nair S, Jao Keehn RJ, Berkebile MM, Maximo JO, Witkowska N, Muller RA (2018) Local resting state functional connectivity in autism: site and cohort variability and the effect of eye status. Brain Imaging Behav 12(1):168–179
25. Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE (2014) Methods to detect, characterize, and remove motion artifact in resting state fmri. Neuroimage 84:320–341
26. Loomes R, Hull L, Mandy WPL (2017) What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. J Am Acad Child Adolesc Psychiatry 56(6):466–474
27. JahediA, Hills T, Olson M, Müller R-A, Fan J. Iterative multivariate matching package for samples with missing data: The iterMatch Package for R (under review)
28. Cox RW (1996) Afni: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29(3):162–173
29. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004) Advances in functional and structural mr image analysis and implementation as fsl. Neuroimage 23(Suppl 1):S208–S219

30. Keown CL, Datko MC, Chen CP, Maximo JO, Jahedi A, Muller RA (2017) Network organization is globally atypical in autism: a graph theory study of intrinsic functional connectivity. Biol Psychiatry Cognit Neurosci Neuroimaging 2(1):66–75
31. Hallquist MN, Hwang K, Luna B (2013) The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fmri preprocessing reintroduces noise and obscures functional connectivity. Neuroimage 82:208–225
32. Gordon EM, Laumann TO, Adeyemo B, Huckins JF, Kelley WM, Petersen SE (2016) Generation and evaluation of a cortical area parcellation from resting-state correlations. Cereb Cortex 26(1):288–303
33. Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31(3):968–980
34. Diedrichsen J, Balsters JH, Flavell J, Cussans E, Ramnani N (2009) A probabilistic mr atlas of the human cerebellum. Neuroimage 46(1):39–46
35. Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Comput Gr Stat 15(3):651–674
36. R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
37. Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2(3):18–22
38. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinform 9:307
39. He Y, Byrge L, Kennedy DP (2020) Nonreplication of functional connectivity differences in autism spectrum disorder across multiple sites and denoising strategies. Hum Brain Mapp 41(5):1334–1350
40. Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, Varoquaux G (2017) Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. Neuroimage 147:736–745
41. Anderson JS, Nielsen JA, Froehlich AL, DuBray MB, Druzgal TJ, Cariello AN, Cooperrider JR, Zielinski BA, Ravichandran C, Fletcher PT, Alexander AL, Bigler ED, Lange N, Lainhart JE (2011) Functional connectivity magnetic resonance imaging classification of autism. Brain 134(Pt 12):3742–3754
42. Kassraian-Fard P, Matthis C, Balsters JH, Maathuis MH, Wenderoth N (2016) Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. Front Psychiatry 7:177
43. Yahata N, Morimoto J, Hashimoto R, Lisi G, Shibata K, Kawakubo Y, Kuwabara H, Kuroda M, Yamada T, Megumi F, Imamizu H, Nanez JE, Takahashi H, Okamoto Y, Kasai K, Kato N, Sasaki Y, Watanabe T, Kawato M (2016) A small number of abnormal brain connections predicts adult autism spectrum disorder. Nat Commun 7:1–2
44. Plitt M, Barnes KA, Martin A (2015) Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. Neuroimage Clin 7:359–366
45. Lombardo MV, Eyler L, Moore A, Datko M, Carter Barnes C, Cha D, Courchesne E, Pierce K (2019) Default mode-visual network hypoconnectivity in an autism subtype with pronounced social visual engagement difficulties. Elife 8:e47427
46. Lynn AC, Padmanabhan A, Simmonds D, Foran W, Hallquist MN, Luna B, O'Hearn K (2018) Functional connectivity differences in autism during face and car recognition: underconnectivity and atypical age-related changes. Dev Sci 21(1):e12508
47. Samson F, Mottron L, Soulieres I, Zeffiro TA (2012) Enhanced visual functioning in autism: an ale meta-analysis. Hum Brain Mapp 33(7):1553–1581
48. Reiter, M. A., Mash, L. E., Linke, A. C., Fong, C. H., Fishman, I., & Muller, R. A. (2018). Distinct Patterns of Atypical Functional Connectivity in Lower-Functioning Autism. Biol PsychiatryCogn Neurosci Neuroimaging, 4(3), 251-259.
49. Farrant K, Uddin LQ (2016) Atypical developmental of dorsal and ventral attention networks in autism. Dev Sci 19(4):550–563
50. Nomi JS, Uddin LQ (2015) Developmental changes in large-scale network connectivity in autism. Neuroimage Clin 7:732–741