Evidence for embracing normative modeling

Saige Rutherford[1,2,3]*, Pieter Barkema[2], Ivy F Tso[3,4], Chandra Sripada[3,5], Christian F Beckmann[1,2,6†], Henricus G Ruhe[2,7†], Andre F Marquand[1,2†]

[1]Department of Cognitive Neuroscience, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands; [2]Donders Institute, Radboud University Nijmegen, Nijmegen, Netherlands; [3]Department of Psychiatry, University of Michigan-Ann Arbor, Ann Arbor, United States; [4]Department of Psychology, University of Michigan-Ann Arbor, Ann Arbor, United States; [5]Department of Philosophy, University of Michigan-Ann Arbor, Ann Arbor, United States; [6]Center for Functional MRI of the Brain (FMRIB), Nuffield Department for Clinical Neuroscience, Welcome Centre for Integrative Neuroimaging, Oxford University, Oxford, United Kingdom; [7]Department of Psychiatry, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands

Abstract In this work, we expand the normative model repository introduced in Rutherford et al., 2022a to include normative models charting lifespan trajectories of structural surface area and brain functional connectivity, measured using two unique resting-state network atlases (Yeo-17 and Smith-10), and an updated online platform for transferring these models to new data sources. We showcase the value of these models with a head-to-head comparison between the features output by normative modeling and raw data features in several benchmarking tasks: mass univariate group difference testing (schizophrenia versus control), classification (schizophrenia versus control), and regression (predicting general cognitive ability). Across all benchmarks, we show the advantage of using normative modeling features, with the strongest statistically significant results demonstrated in the group difference testing and classification tasks. We intend for these accessible resources to facilitate the wider adoption of normative modeling across the neuroimaging community.

Editor's evaluation

This is a rigorous and compelling extension of previous normative modeling work. The current study demonstrates that normative models incorporating lifespan trajectories of structural and functional connectivity provide a strong basis for brain imaging studies across a range of tasks including, univariate group difference assessment, classification, and building regression models. The work is important, rigorous and a valuable contribution to the field.

***For correspondence:** saige.rutherford@donders.ru.nl

†These authors contributed equally to this work

Competing interest: See page 14

Funding: See page 14

Preprinted: 15 November 2022
Received: 28 November 2022
Accepted: 10 March 2023
Published: 13 March 2023

Reviewing Editor: Chris I Baker, National Institute of Mental Health, United States

Introduction

Normative modeling is a framework for mapping population-level trajectories of the relationships between health-related variables while simultaneously preserving individual-level information (***Marquand et al., 2016a***; ***Marquand et al., 2016b***; ***Rutherford et al., 2022b***). Health-related variables is an intentionally inclusive and broad definition that may involve demographics (i.e. age and gender), simple (i.e. height and weight), or complex (i.e. brain structure and function, genetics) biological measures, environmental factors (i.e. urbanicity, pollution), self-report measures (i.e. social satisfaction, emotional experiences), or behavioral tests (i.e. cognitive ability, spatial reasoning). Charting the relationships, as mappings between a covariate (e.g. age) and response variable (e.g.

brain measure) in a reference population creates a coordinate system that defines the units in which humans vary. Placing individuals into this coordinate system creates the opportunity to characterize their profiles of deviation. While this is an important aspect of normative modeling, it is usually just the first step, i.e., you are often interested in using the outputs of normative models in downstream analyses to detect case-control differences, stratification, or individual statistics. This framework provides a platform for such analyses as it effectively translates diverse data to a consistent scale, defined with respect to population norms.

Normative modeling has seen widespread use spanning diverse disciplines. The most well-known example can be found in pediatric medicine, where conventional growth charts are used to map the height, weight, and head circumference trajectories of children (***Borghi et al., 2006***). Under the neuroscience umbrella, generalizations of this approach have been applied in the fields of psychiatry (***Floris et al., 2021***; ***Madre et al., 2020***; ***Wolfers et al., 2015***; ***Wolfers et al., 2017***; ***Wolfers et al., 2018***; ***Wolfers et al., 2020***; ***Wolfers et al., 2021***; ***Zabihi et al., 2019***; ***Zabihi et al., 2020***), neurology (***Itälinna et al., 2022***; ***Verdi et al., 2021***), developmental psychology (***Holz et al., 2022***; ***Kjelkenes et al., 2022***), and cognitive neuroscience (***Marquand et al., 2017***). Throughout these numerous applications, normative models have exposed the shortcomings of prior case-control frameworks, i.e., that they rely heavily on the assumption, there is within-group homogeneity. This case versus control assumption is often an oversimplification, particularly in psychiatric diagnostic categories, where the clinical labels used to place individuals into group categories are often unreliable, poorly measured, and may not map cleanly onto underlying biological mechanisms (***Cai et al., 2020***; ***Cuthbert and Insel, 2013***; ***Flake and Fried, 2020***; ***Insel et al., 2010***; ***Linden, 2012***; ***Loth et al., 2021***; ***Michelini et al., 2021***; ***Moriarity et al., 2022***; ***Moriarity and Alloy, 2021***; ***Nour et al., 2022***; ***Sanislow, 2020***; ***Zhang et al., 2021***). Correspondingly, traditional analysis techniques for modeling case versus control effects have often led to null findings (***Winter et al., 2022***) or significant but very small clinically meaningless differences. These effects are furthermore frequently aspecific to an illness or disorder (***Baker et al., 2019***; ***Goodkind et al., 2015***; ***McTeague et al., 2017***; ***Sprooten et al., 2017***) and inconsistent or contradictory (***Filip et al., 2022***; ***Lee et al., 2007***; ***Pereira-Sanchez and Castellanos, 2021***) yielding questionable clinical utility (***Etkin, 2019***; ***Mottron and Bzdok, 2022***).

In addition to the applications of normative modeling, there is also active technical development (***Dinga et al., 2021***; ***Fraza et al., 2022***; ***Fraza et al., 2021***; ***Kia et al., 2020***; ***Kia et al., 2021***; ***Kia and Marquand, 2018***; ***Kumar, 2021***). Due to the growing popularity of normative modeling and in recognition of the interdisciplinary requirements using and developing this technology (clinical domain knowledge, statistical expertise, data management, and computational demands), research interests have been centered on open science, and inclusive, values (***Gau et al., 2021***; ***Levitis et al., 2021***) that support this type of interdisciplinary scientific work. These values encompass open-source software, sharing pre-trained big data models (***Rutherford et al., 2022a***), online platforms for communication and collaboration, extensive documentation, code tutorials, and protocol-style publications (***Rutherford et al., 2022b***).

The central contribution of this paper is to, first, augment the models in ***Rutherford et al., 2022a***, with additional normative models for surface area and functional connectivity, which are made open and accessible to the community. Second, we comprehensively evaluate the utility of normative models for a range of downstream analyses, including (1) mass univariate group difference testing (schizophrenia versus controls), (2) multivariate prediction – classification (using support vector machines to distinguish schizophrenia from controls), and (3) multivariate prediction – regression (using principal component regression (PCR) to predict general cognitive ability) (***Figure 1***). Within these benchmarking tasks, we show the benefit of using normative modeling features compared to using raw features. We aim for these benchmarking results, along with our publicly available resources (code, documentation, tutorials, protocols, community forum, and website for running models without using any code). Combined this provides practical utility as well as scientific evidence for embracing normative modeling.

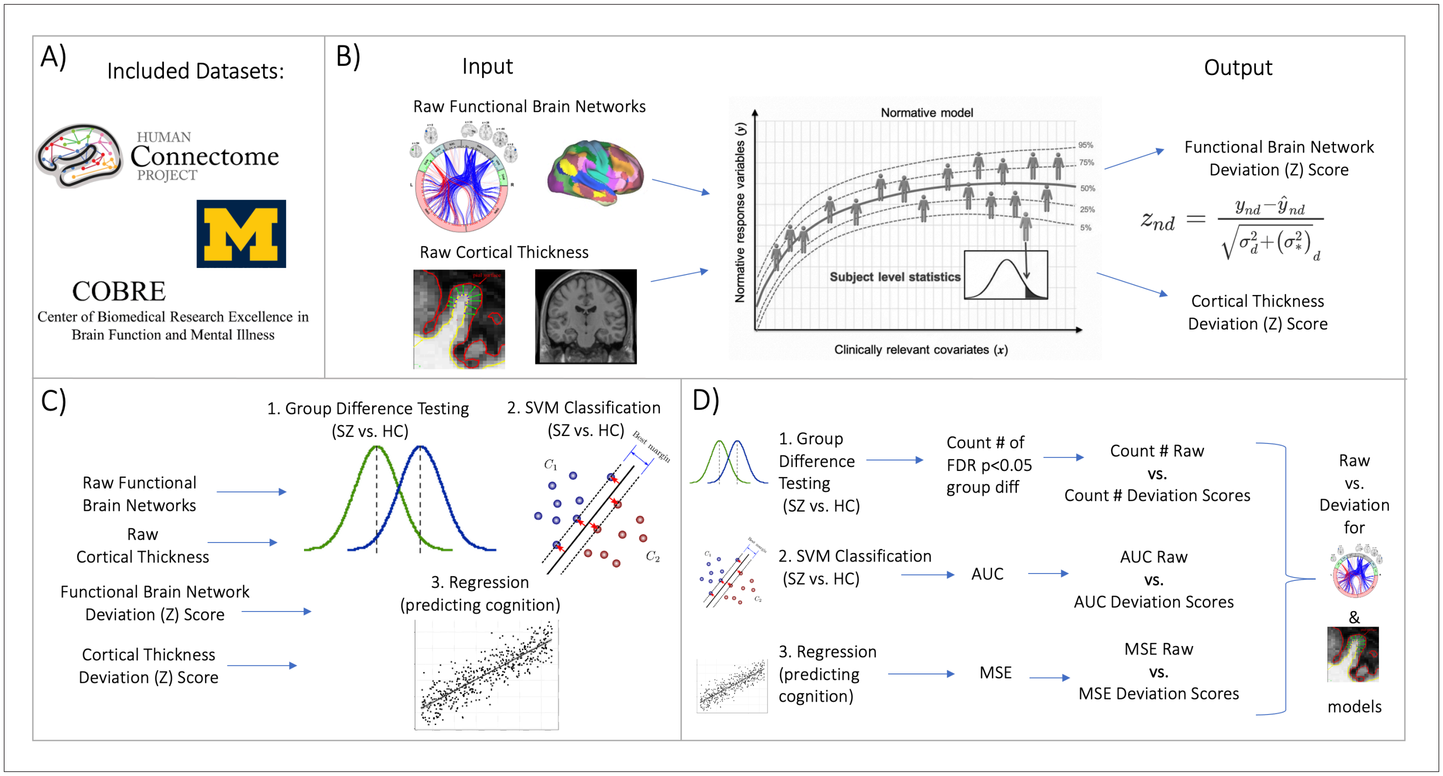

A)
Included Datasets:
HUMAN Connectome PROJECT
COBRE
Center of Biomedical Research Excellence in Brain Function and Mental Illness
B)
Input
Raw Functional Brain Networks
Raw Cortical Thickness
Normative model
Normative response variables (y)
Clinically relevant covariates (x)
Subject level statistics
95%
75%
50%
25%
5%
Output
Functional Brain Network Deviation (Z) Score
$z_{nd} = \frac{y_{nd} - \hat{y}_{nd}}{\sqrt{\sigma_d^2 + (\sigma_*^2)_d}}$
Cortical Thickness Deviation (Z) Score
C)
1. Group Difference Testing (SZ vs. HC)
2. SVM Classification (SZ vs. HC)
Best margin
C_1
C_2
Raw Functional Brain Networks
Raw Cortical Thickness
Functional Brain Network Deviation (Z) Score
Cortical Thickness Deviation (Z) Score
3. Regression (predicting cognition)
D)
1. Group Difference Testing (SZ vs. HC)
Count # of FDR p<0.05 group diff
Count # Raw vs. Count # Deviation Scores
2. SVM Classification (SZ vs. HC)
AUC
AUC Raw vs. AUC Deviation Scores
3. Regression (predicting cognition)
MSE
MSE Raw vs. MSE Deviation Scores
Raw vs. Deviation for
&
models

Figure 1. Overview of workflow. (**A**) Datasets included the Human Connectome Project (young adult) study, the University of Michigan schizophrenia study, and the Center for Biomedical Research Excellence (COBRE) schizophrenia study. (**B**) Openly shared, pre-trained on big data, normative models were estimated for large-scale resting-state functional brain networks and cortical thickness. (**C**) Deviation (Z) scores and raw data, for both functional and structural data, were input into three benchmarking tasks: 1. group difference testing, 2. support vector machine (SVM) classification, and 3. regression (predicting cognition). (**D**) Evaluation metrics were calculated for each benchmarking task. These metrics were calculated for the raw data models and the deviation score models. The difference between each models' performance was calculated for both functional and structural modalities.

Methods

Dataset selection and scanner parameters

Datasets used for training the functional normative models closely match the sample included in ***Rutherford et al., 2022a***, apart from sites that did not collect or were unable to share functional data. Evaluation of the functional normative models was performed in a test set (20% of the training set) and in two transfer sets that are comprised of scanning sites not seen by the model during training (clinical and healthy controls). The full details of the data included in the functional normative model training can be found in Appendix 1 and ***Supplementary file 1***. We leverage several datasets for the benchmarking tasks, the Human Connectome Project Young Adult study (HCP) (***Van Essen et al., 2013***), The Center for Biomedical Research Excellence (COBRE) (***Aine et al., 2017***; ***Sui et al., 2018***), and the University of Michigan SchizGaze (UMich) (***Tso et al., 2021***; ***Table 1***). The HCP data was chosen because it is widely used by the neuroscience community, especially for prediction studies. Also, prior studies using HCP data have shown promising results for predicting general cognitive ability (***Sripada***

Table 1. Dataset inclusion and sample overview.

		Cortical Thickness			Functional Networks		
Study	Benchmark Task	N	Age (m, s.d.)	F, M (%)	N	Age (m, s.d.)	F, M (%)
HCP	Regression – predicting cognition	529	28.8, 3.6	53.4, 46.6	499	28.9, 3.6	54.3, 45.6
COBRE	Classification & Group Difference	124	37.0, 12.7	24.2, 75.8	121	35.4, 12.4	23.1, 76.9
UMich	Classification & Group Difference	89	32.6, 9.6	50.6, 49.3	87	33.0, 10.1	50.6, 49.3

et al., 2020a). The HCP data was used in the prediction – regression benchmarking task. The COBRE and UMich datasets are used in the classification and group difference testing benchmarking tasks. Inclusion criteria across all the datasets were that the participant has necessary behavioral and demographic variables, as well as high-quality MRI data. High-quality was defined for structural images as in our prior work (***Rutherford et al., 2022a***), namely as the lack of any artifacts such as ghosting or ringing, that Freesurfer surface reconstruction was able to run successfully, and that the Euler number calculated from Freesurfer (***Klapwijk et al., 2019***), which is a proxy metric for scan quality, was below a chosen threshold (rescaled Euler <10) (***Kia et al., 2022***). High-quality functional data followed recommended practices (***Siegel et al., 2017***) and was defined as having a high-quality structural MRI (required for co-registration and normalization) and at least 5 min of low motion data (framewise displacement <0.5 mm). The HCP, COBRE, and UMich functional and structural data were manually inspected for quality at several tasks during preprocessing (after co-registration of functional and structural data and after normalization of functional data to MNI template space).

All subjects provided informed consent. Subject recruitment procedures and informed consent forms, including consent to share de-identified data, were approved by the corresponding university institutional review board where data were collected. The scanning acquisition parameters were similar but varied slightly across the studies, details in Appendix 1.

Demographic, cognition, and clinical diagnosis variables

Demographic variables included age, sex, and MRI scanner site. A latent variable of cognition, referred to as General Cognitive Ability (GCA), was created for the regression benchmarking task using HCP data. The HCP study administered the NIHToolbox Cognition battery (***Gershon et al., 2010***), and a bi-factor model was fit (for further modeling details and assessment of model fit see ***Sripada et al., 2020b***). For COBRE and UMich studies, clinical diagnosis of schizophrenia was confirmed using the Structured Clinical Interview used for DSM-5 disorders (SCID) (***First, 1956***). All subjects were screened and excluded if they had: a history of neurological disorder, mental retardation, severe head trauma, or substance abuse/dependence within the last 6 (UMich) or 12 months (COBRE), were pregnant/nursing (UMich), or had any contraindications for MRI.

Image preprocessing

Structural MRI data were preprocessed using the Freesurfer (version 6.0) recon-all pipeline (***Dale et al., 1999***; ***Fischl et al., 2002***; ***Fischl and Dale, 2000***) to reconstruct surface representations of the volumetric data. Estimates of cortical thickness and subcortical volume were then extracted (aparc and aseg) for each subject from their Freesurfer output folder, then merged, and formatted into a csv file (rows = subjects, columns = brain ROIs). We also share models of surface area, extracted in the same manner as the cortical thickness data from a similar dataset (described in ***Supplementary file 2***).

Resting-state data were preprocessed separately for each study using fMRIPrep ***Esteban et al., 2019***; however, similar steps were done to all resting-state data following best practices including field-map correction of multi-band data, slice time correction (non-multi-band data), co-registration of functional to structural data, normalization to MNI template space, spatial smoothing (2 x voxel size, 4–6 mm), and regression of nuisance confounders (WM/CSF signals, non-aggressive AROMA components [***Pruim et al., 2015a***; ***Pruim et al., 2015b***], linear and quadratic effects of motion).

Large-scale brain networks from the 17-network Yeo atlas (***Yeo et al., 2011***) were then extracted and between-network connectivity was calculated using full correlation. We also shared functional normative models using the Smith-10 ICA-based parcellation (***Smith et al., 2009***) which includes subcortical coverage, however, the benchmarking tasks only use the Yeo-17 functional data. Fisher r-to-z transformation was performed on the correlation matrices. If there were multiple functional runs, connectivity matrices were calculated separately for each run then all runs for a subject were averaged. For further details regarding the preparation of the functional MRI data, see Appendix 1.

Normative model formulation

After dataset selection and preprocessing, normative models were estimated using the Predictive Clinical Neuroscience toolkit (PCNtoolkit), an open-source python package for normative modeling (***Marquand et al., 2021***). For the structural data, we used a publicly shared repository of pre-trained normative models that were estimated on approximately 58,000 subjects using a warped Bayesian

Linear Regression algorithm (***Fraza et al., 2021***). The covariates used to train the structural normative models included age, sex, data quality metric (Euler number), and site. Normative models of surface area were also added to the same repository ***Supplementary file 2***. Model fit was established using explained variance, mean standardized log loss, skew, and kurtosis. The outputs of normative modeling also include a Z-score, or deviation score, for all brain regions and all subjects. The deviation score represents where the individual is in comparison to the population the model was estimated on, where a positive deviation score corresponds to the greater cortical thickness or subcortical volume than average, and a negative deviation score represents less cortical thickness or subcortical volume than average. The deviation (Z) scores that are output from the normative model are the features input for the normative modeling data in the benchmarking analyses.

In addition to normative models of brain structure, we also expanded our repository by estimating normative models of brain functional connectivity (resting-state brain networks, Yeo-17 and Smith-10) using the same algorithm (Bayesian Linear Regression) as the structural models. The covariates used to train the functional normative models were similar to the structural normative models which included age, sex, data quality metric (mean framewise displacement), and site. Functional normative models were trained on a large multi-site dataset (approx. N=22,000) and evaluated in several test sets using explained variance, mean standardized log loss, skew, and kurtosis. The training dataset excluded subjects with any known psychiatric diagnosis. We transferred the functional normative models to the datasets used in this work for benchmarking (***Table 1***) to generate deviation (Z) scores. HCP was included in the initial training (half of the sample was held out in the test set), while the UMich and COBRE datasets were not included in the training and can be considered as examples of transfer to new, unseen sites.

"Raw" input data

The data that we compare the output of normative modeling to, referred to throughout this work as 'raw' input data, is simply the outputs of traditional preprocessing methods for structural and functional MRI. For structural MRI, this corresponds to the cortical thickness files that are output after running the Freesurfer recon-all pipeline. We used the aparcstats2table and asegstats2table functions to extract the cortical thickness and subcortical volume from each region in the Destrieux atlas and Freesurfer subcortical atlas. For functional MRI, tradition data refers to the Yeo17 brain network connectomes which were extracted from the normalized, smoothed, de-noised functional time-series. The upper triangle of each subject's symmetric connectivity matrix was vectorized, where each cell represents a unique between-network connection. For clarification, we also note that the raw input data is the starting point of the normative modeling analysis, or in other words, the raw input data is the response variable or independent (Y) variable that is predicted from the vector of covariates when estimating the normative model. Before entering into the benchmarking tasks, to create a fair comparison between raw data and deviation scores, nuisance variables including sex, site, linear and quadratic effects of age and head motion (only for functional models) were regressed out of the raw data (structural and functional) using least squares regression.

Benchmarking

The benchmarking was performed in three separate tasks, mass univariate group difference testing, multivariate prediction – classification, and multivariate prediction – regression, described in further detail below. In each benchmarking task, a model was estimated using the deviation scores as input features and then estimated again using the raw data as the input features. For task one, group difference testing, the models fit in a univariate approach meaning there was one test performed for each brain feature, and for tasks 2 and 3, classification and regression, the models fit in a multivariate approach. After each model was fit, the performance metrics were evaluated and the difference in performance between the deviation score and raw data models was calculated, again described in more detail below.

Task one: Mass univariate group difference testing

Mass univariate group difference (schizophrenia versus control) testing was performed across all brain regions. Two sample independent t-tests were estimated and run on the data using the SciPy python package (***Virtanen et al., 2020***). After addressing multiple comparison corrections, brain regions with

FDR corrected p<.05 were considered significant and the total number of regions displaying statistically significant group differences was counted.

For the purpose of comparing group difference effects to individual differences, we also summarized the individual deviation maps and compare this map to the group difference map. Individual deviation maps were summarized by counting the number of individuals with 'extreme' deviations (Z>2 or Z<–2) at a given brain region or network connectivity pair. This was done separately for positive and negative deviations and for each group and visualized qualitatively (Figure 4B). To quantify the individual difference maps in comparison to group differences, we performed a Mann-Whitney U-test on the count of extreme deviations in each group. The U-test was used because the distribution of count data is skewed (non-Gaussian) which the U-test is designed to account for (***Mann and Whitney, 1947***).

Task two: Multivariate prediction – classification

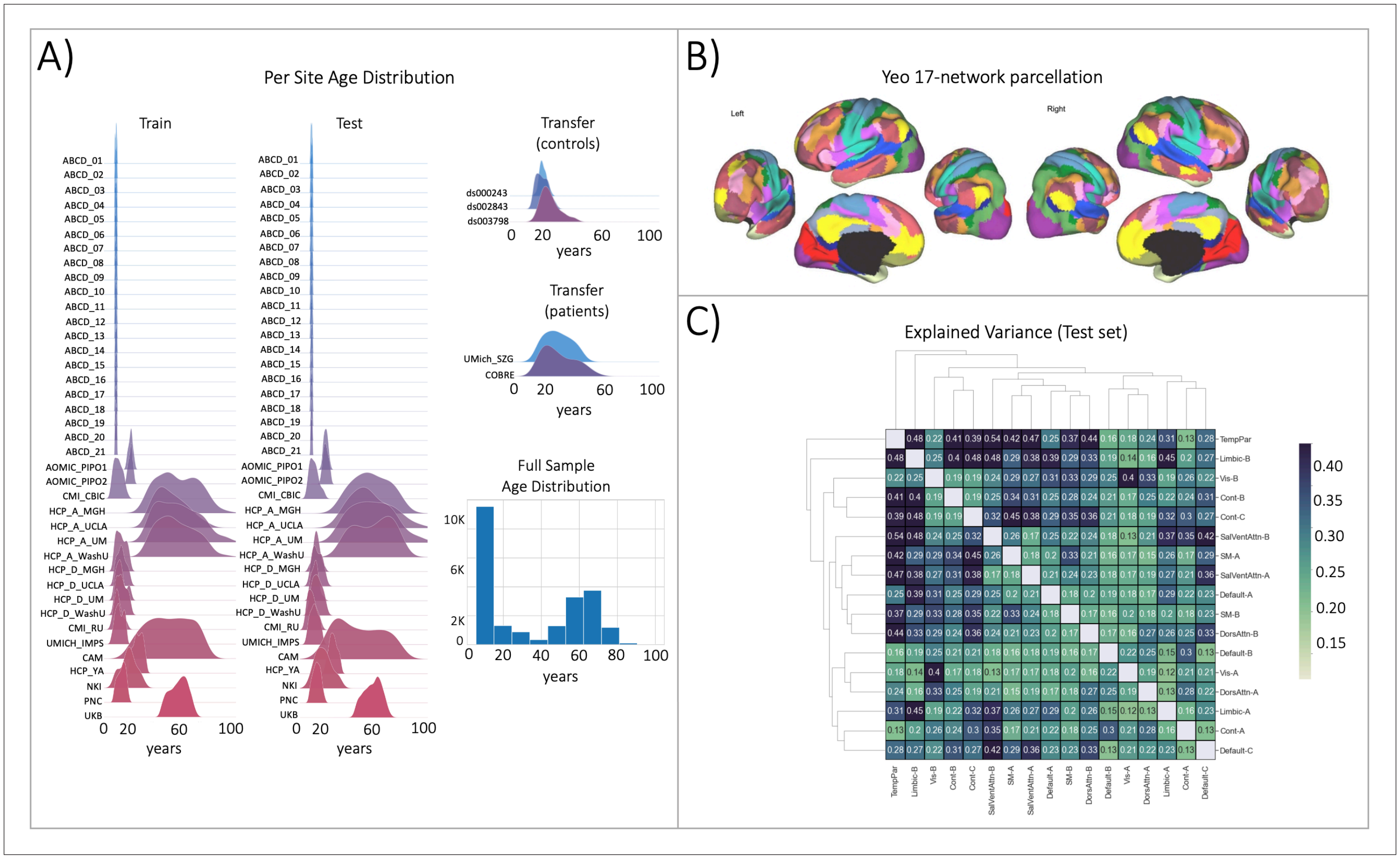
Support vector machine is a commonly used algorithm in machine learning studies and performs well in classification settings. A support vector machine constructs a set of hyper-planes in a high dimensional space and optimizes to find the hyper-plane that has the largest distance, or margin, to the nearest training data points of any class. A larger margin represents better linear separation between classes and will correspond to a lower error of the classifier in new samples. Samples that lie on the margin boundaries are also called 'support vectors.' The decision function provides per-class scores than can be turned into probabilities estimates of class membership. We used Support vector classification (SVC) with a linear kernel as implemented in the scikit-learn package (version 1.0.9) (***Pedregosa et al., 2011***) to classify a schizophrenia group from a control group. These default hyperparameters were chosen based on following an example of SVC provided by scikit-learn, however, similar results were obtained using a radial basis function kernel (not shown). This classification setting of distinguishing schizophrenia from a control group was chosen due to past work showing the presence of both case-control group differences and individual differences (***Wolfers et al., 2018***). The evaluation metric for the classification task is an area under the receiving operator curve (AUC) averaged across all folds within a 10-fold cross-validation framework.

Task three: Multivariate prediction – regression

A linear regression model was implemented to predict a latent variable of cognition (general cognitive ability) in the HCP dataset. Brain Basis Set (BBS) is a predictive modeling approach developed and validated in previous studies (***Sripada et al., 2019***; ***Sripada et al., 2019***); see also studies by Wager and colleagues for a broadly similar approach (***Chang et al., 2015***; ***Wager et al., 2013***; ***Woo et al., 2017***). BBS is similar to principal component regression (***Jolliffe, 1982***; ***Park, 1981***), with an added predictive element. In the training set, PCA is performed on a *n*_subjects by *p*_brain_features matrix using the PCA function from scikit-learn in Python, yielding components ordered by descending eigenvalues. Expression scores are then calculated for each of the *k* components for each subject by projecting each subject's feature matrix onto each component. A linear regression model is then fit with these expression scores as predictors and the phenotype of interest (general cognitive ability) as the outcome, saving **B**, the *k x 1* vector of fitted coefficients, for later use. In a test partition, the expression scores for each of the *k* components for each subject are again calculated. The predicted phenotype for each test subject is the dot product of **B** learned from the training partition with the vector of component expression scores for that subject. We set k=15 in all models, following prior work (***Rutherford et al., 2020***). The evaluation metric for the regression task is the mean squared error of the prediction in the test set.

Benchmarking: Model comparison evaluation

Evaluation metrics of each task (count, AUC, and MSE) were calculated independently for both deviation score (Z) and raw data (R) models. Higher AUC, higher count, and lower MSE represent better



	TempPar	Limbic-B	Vis-B	Cont-B	Cont-C	SalVentAttn-B	SM-A	SalVentAttn-A	Default-A	SM-B	DorsAttn-B	Default-B	Vis-A	DorsAttn-A	Limbic-A	Cont-A	Default-C
TempPar		0.48	0.22	0.41	0.39	0.54	0.42	0.47	0.25	0.37	0.44	0.16	0.18	0.24	0.31	0.13	0.28
Limbic-B	0.48		0.25	0.4	0.48	0.48	0.29	0.38	0.39	0.29	0.33	0.19	0.14	0.16	0.45	0.2	0.27
Vis-B	0.22	0.25		0.19	0.19	0.24	0.29	0.27	0.31	0.33	0.29	0.25	0.4	0.33	0.19	0.26	0.22
Cont-B	0.41	0.4	0.19		0.19	0.25	0.34	0.31	0.25	0.28	0.24	0.21	0.17	0.25	0.22	0.24	0.31
Cont-C	0.39	0.48	0.19	0.19		0.32	0.45	0.38	0.29	0.35	0.36	0.21	0.18	0.19	0.32	0.3	0.27
SalVentAttn-B	0.54	0.48	0.24	0.25	0.32		0.26	0.17	0.25	0.22	0.24	0.18	0.13	0.21	0.37	0.35	0.42
SM-A	0.42	0.29	0.29	0.34	0.45	0.26		0.18	0.2	0.33	0.21	0.16	0.17	0.15	0.26	0.17	0.29
SalVentAttn-A	0.47	0.38	0.27	0.31	0.38	0.17	0.18		0.21	0.24	0.23	0.18	0.17	0.19	0.27	0.21	0.36
Default-A	0.25	0.39	0.31	0.25	0.29	0.25	0.2	0.21		0.18	0.2	0.19	0.18	0.17	0.29	0.22	0.23
SM-B	0.37	0.29	0.33	0.28	0.35	0.22	0.33	0.24	0.18		0.17	0.16	0.2	0.18	0.2	0.18	0.23
DorsAttn-B	0.44	0.33	0.29	0.24	0.36	0.24	0.21	0.23	0.2	0.17		0.17	0.16	0.27	0.26	0.25	0.33
Default-B	0.16	0.19	0.25	0.21	0.21	0.18	0.16	0.18	0.19	0.16	0.17		0.22	0.25	0.15	0.3	0.13
Vis-A	0.18	0.14	0.4	0.17	0.18	0.13	0.17	0.17	0.18	0.2	0.16	0.22		0.19	0.12	0.21	0.21
DorsAttn-A	0.24	0.16	0.33	0.25	0.19	0.21	0.15	0.19	0.17	0.18	0.27	0.25	0.19		0.13	0.28	0.22
Limbic-A	0.31	0.45	0.19	0.22	0.32	0.37	0.26	0.27	0.29	0.2	0.26	0.15	0.12	0.13		0.16	0.23
Cont-A	0.13	0.2	0.26	0.24	0.3	0.35	0.17	0.21	0.22	0.18	0.25	0.3	0.21	0.28	0.16		0.13
Default-C	0.28	0.27	0.22	0.31	0.27	0.42	0.29	0.36	0.23	0.23	0.33	0.13	0.21	0.22	0.23	0.13	

Figure 2. Functional brain network normative modeling. (**A**) Age distribution per scanning site in the train, test, and transfer data partitions and across the full sample (train +test). (**B**) The Yeo-17 brain network atlas is used to generate connectomes. Between network connectivity was calculated for all 17 networks, resulting in 136 unique network pairs that were each individually input into a functional normative model. (**C**) The explained variance in the controls test set (N=7244) of each of the unique 136 network pairs of the Yeo-17 atlas. Networks were clustered for visualization to show similar variance patterns.

model performance. We then have a statistic of interest that is observed, theta, which represents the difference between deviation and raw data model performance.

$$\theta_{task\ 1} = Count_z - \ Count_R$$
$$\theta_{task\ 2} = AUC_z - AUC_R$$
$$\theta_{task\ 3} = \ MSE_R - MSE_z$$

To assess whether θ is more likely than would be expected by chance, we generated the null distribution for theta using permutations. Within one iteration of the permutation framework, a random sample is generated by shuffling the labels (In tasks 1 & 2 we shuffle SZ/HC labels, and in task three we shuffle cognition labels). Then this sample is used to train both deviation and raw models, ensuring the same row shuffling scheme across both deviation score and raw data datasets (for each permutation iteration). The shuffled models are evaluated, and we calculate θ_{perm} for each random shuffle of labels. We set n_permutations =10,000 and use the distribution of θ_{perm} to calculate a p-value for $\theta_{observed}$ at each benchmarking task. The permuted p-value is equal to $(\mathrm{C} + 1)/(\mathrm{n\_permutations} + 1)$. Where C is the number of permutations where $\theta_{perm} >= \theta_{observed}$. The same evaluation procedure described here 293 (including permutations) was performed for both cortical thickness and functional network modalities.

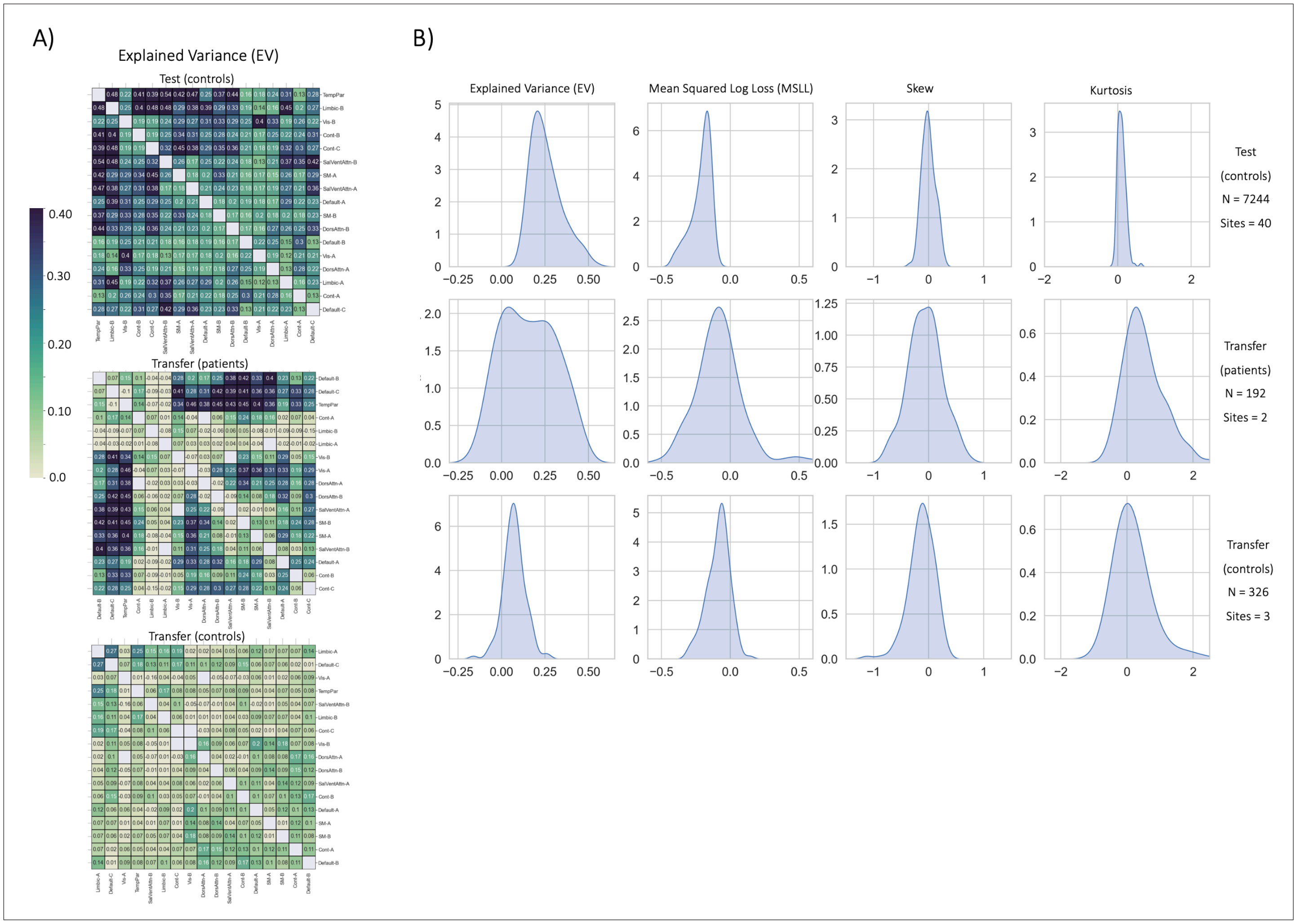


Figure 3. Functional normative model evaluation metrics. (**A**) Explained variance per network pair across the test set (top), and both transfer sets (patients – middle, controls – bottom). Networks were clustered for visualization to show similar variance patterns. (**B**) The distribution across all models of the evaluation metrics (columns) in the test set (top row) and both transfer sets (middle and bottom rows). Higher explained variance (closer to one), more negative MSLL, and normally distributed skew and kurtosis correspond to better model fit.

Results

Sharing of functional big data normative models

The first result of this work is the evaluation of the functional big data normative models (***Figure 2***). These models build upon the work of ***Rutherford et al., 2022a*** in which we shared population-level structural normative models charting cortical thickness and subcortical volume across the human lifespan (ages 2–100). The datasets used for training the functional models, the age range of the sample, and the procedures for evaluation closely resemble the structural normative models. The sample size (approx. N=22,000) used for training and testing the functional models is smaller than the structural models (approx. N=58,000) due to data availability (i.e. some sites included in the structural models did not collect functional data or could not share the data) and the quality control procedures (see methods). However, despite the smaller sample size of the functional data reference cohort, the ranges of the evaluation metrics are quite similar to the structural models (***Figure 3***). Most importantly, we demonstrate the opportunity to transfer the functional models to new samples, or sites that were not included in the original training and testing sets, referred to as the transfer set, and show that transfer works well in a clinical sample (***Figure 3*** - transfer patients) or sample of healthy controls (***Figure 3*** - transfer controls).

Table 2. Benchmarking results.

Deviation (Z) score column shows the performance using deviation scores (AUC for classification, the total number of regions with significant group differences FDR-corrected p<0.05 for case versus control, mean squared error for regression), Raw column represents the performance when using the raw data, and Difference column shows the difference between the deviation scores and raw data (Deviation - Raw). Higher AUC, higher count, and lower MSE represent better performance. Positive values in the Difference column show that there is better performance when using deviation scores as input features for classification and group difference tasks, and negative performance difference values for the regression task show there is a better performance using the deviation scores. *=statistically significant difference between Z and Raw established using permutation testing (10 k perms).

Benchmark	Modality	Normative Modeling Deviation Score Data	Raw Data	Performance Difference
Group Difference	Cortical thickness	117/187	0/187	117*
Group Difference	Functional Networks	50/136	0/136	50*
Classification	Cortical thickness	0.87	0.43	0.44*
Classification	Functional Networks	0.69	0.68	0.01
Regression	Cortical thickness	0.699	0.708	–0.008
Regression	Functional Networks	0.877	0.890	–0.013

Normative modeling shows larger effect sizes in mass univariate group differences

The strongest evidence for embracing normative modeling can be seen in the benchmarking task one group difference (schizophrenia versus controls) testing results (***Table 2***, ***Figure 4***). In this application, we observe numerous group differences in both functional and structural deviation score models after applying stringent multiple comparison corrections (FDR p-value <0.05). The strongest effects (HC>SZ) in the structural models were located in the right hemisphere lateral occipitotemporal sulcus (S_oc_temp_lat) thickness, right hemisphere superior segment of the circular sulcus of the insula (S_circular_ins_sup) thickness, right Accumbens volume, left hemisphere Supramarginal gyrus (G_pariet_inf_Supramar) thickness, and left hemisphere Inferior occipital gyrus (O3) and sulcus (G_and_S_occipital_inf) thickness. For the functional models, the strongest effects (HC>SZ t-statistic) were observed in the between-network connectivity of Visual A-Default B, Dorsal Attention A-Control B, and Visual B-Limbic A. In the raw data models, which were residualized of covariates including site, sex, and linear +quadratic effects of age and head motion (only included for functional models), we observe no group differences after multiple comparison corrections. The lack of any group differences in the raw data was initially a puzzling finding due to reported group differences in the literature (***Arbabshirani et al., 2013***; ***Cetin et al., 2015***; ***Cetin et al., 2016***; ***Cheon et al., 2022***; ***Dansereau et al., 2017***; ***Howes et al., 2023***; ***Lei et al., 2020a***; ***Lei et al., 2020b***; ***Meng et al., 2017***; ***Rahim et al., 2017***; ***Rosa et al., 2015***; ***Salvador et al., 2017***; ***Shi et al., 2021***; ***van Erp et al., 2018***; ***Venkataraman et al., 2012***; ***Wannan et al., 2019***; ***Yu et al., 2012***), however, upon the investigation of the uncorrected statistical maps, we observe that the raw data follows a similar pattern to the deviation group difference map (***Figure 4***), but these results do not withstand multiple comparison correction. For full statistics including the corrected and uncorrected p-values and test-statistic of every ROI, see ***Supplementary files 3 and 4***. While there have been reported group differences between controls and schizophrenia in cortical thickness and resting state brain networks in the literature, these studies have used different datasets (of varying sample sizes), different preprocessing pipelines and software versions, and different statistical frameworks (***Castro et al., 2016***; ***Di Biase et al., 2019***; ***Dwyer et al., 2018***; ***Geisler et al., 2015***; ***Marek et al., 2022***; ***Sui et al., 2015***; ***van Haren et al., 2011***). When reviewing the literature of studies on SZ versus HC group difference testing, we did not find any study that performed univariate t-testing and multiple comparison correction at the ROI-level or network-level, rather most works used statistical tests and multiple comparison correction at the voxel-level or edge-level. Combined with the known patterns of heterogeneity present in schizophrenia disorder (***Lv et al., 2021***; ***Wolfers et al., 2018***), it is unsurprising that our results differ from past studies.

The qualitative (***Figure 4B***) and quantitative (***Figure 4C***) comparison of the group difference maps with the individual difference maps showed the additional benefit of normative modeling - that it can

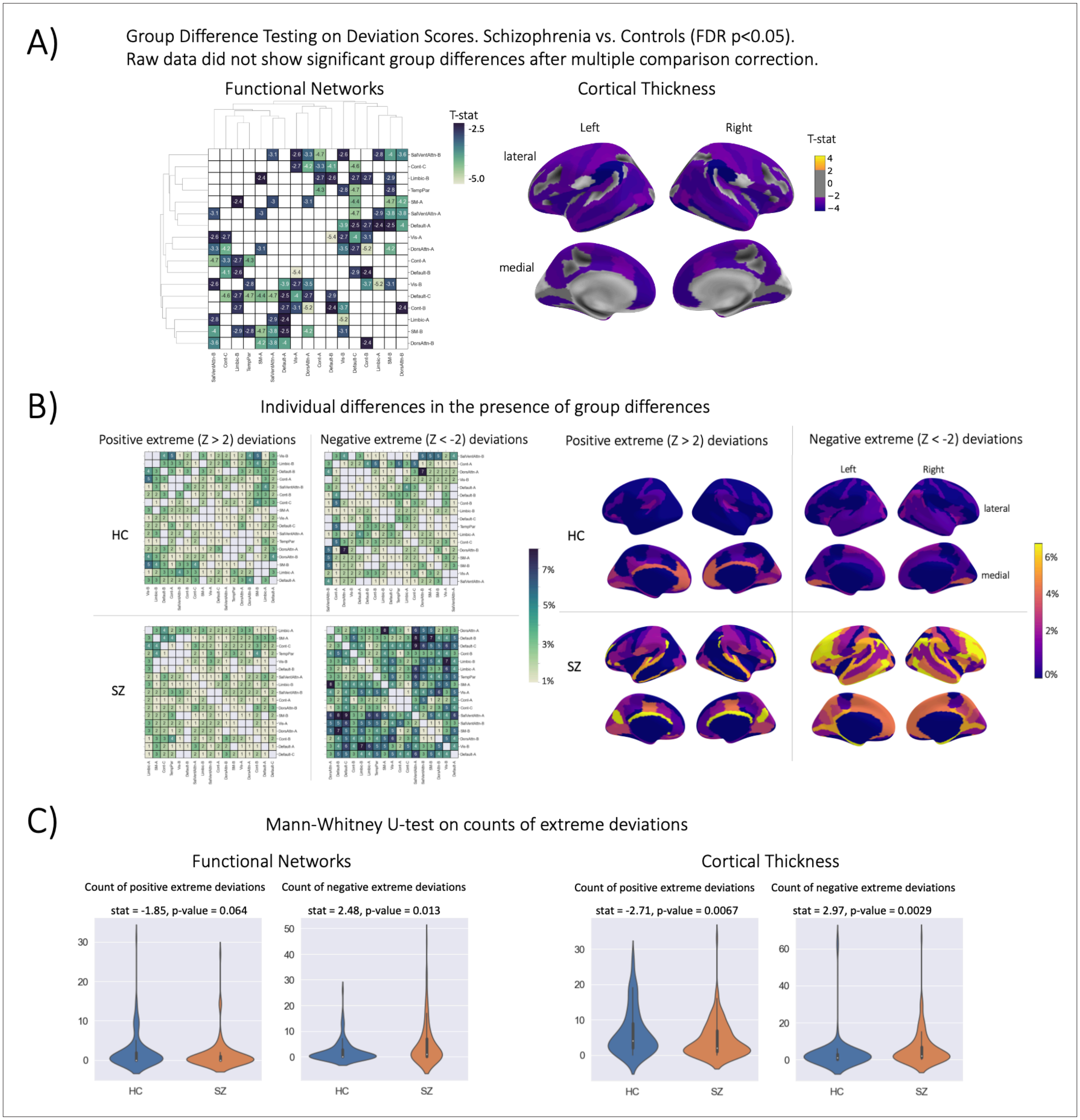


Figure 4. Group difference testing evaluation. (**A**) Significant group differences in the deviation score models, (top left) functional brain network deviation, and (top right) cortical thickness deviation scores. The raw data, either cortical thickness or functional brain networks (residualized of sex and linear/ quadratic effects of age and motion (mean framewise displacement)) resulted in no significant group differences after multiple comparison corrections. Functional networks were clustered for visualization to show similar variance patterns. (**B**) There are still individual differences observed that do not overlap with the group difference map, showing the benefit of normative modeling, which can detect both group and individual differences through proper modeling of variation. Functional networks were clustered for visualization to show similar variance patterns. (**C**) There are significant group differences in the summaries (count) of the individual difference maps (panel B).

reveal subtle individual differences which are lost when only looking at group means. The individual difference maps show that at every brain region or connection, there is at least one person, across both patient and clinical groups, that has an extreme deviation. We found significant differences in the count of negative deviations (SZ >HC) for both cortical thickness (p=0.0029) and functional networks (p=0.013), and significant differences (HC >SZ) in the count of positive cortical thickness (p=0.0067).

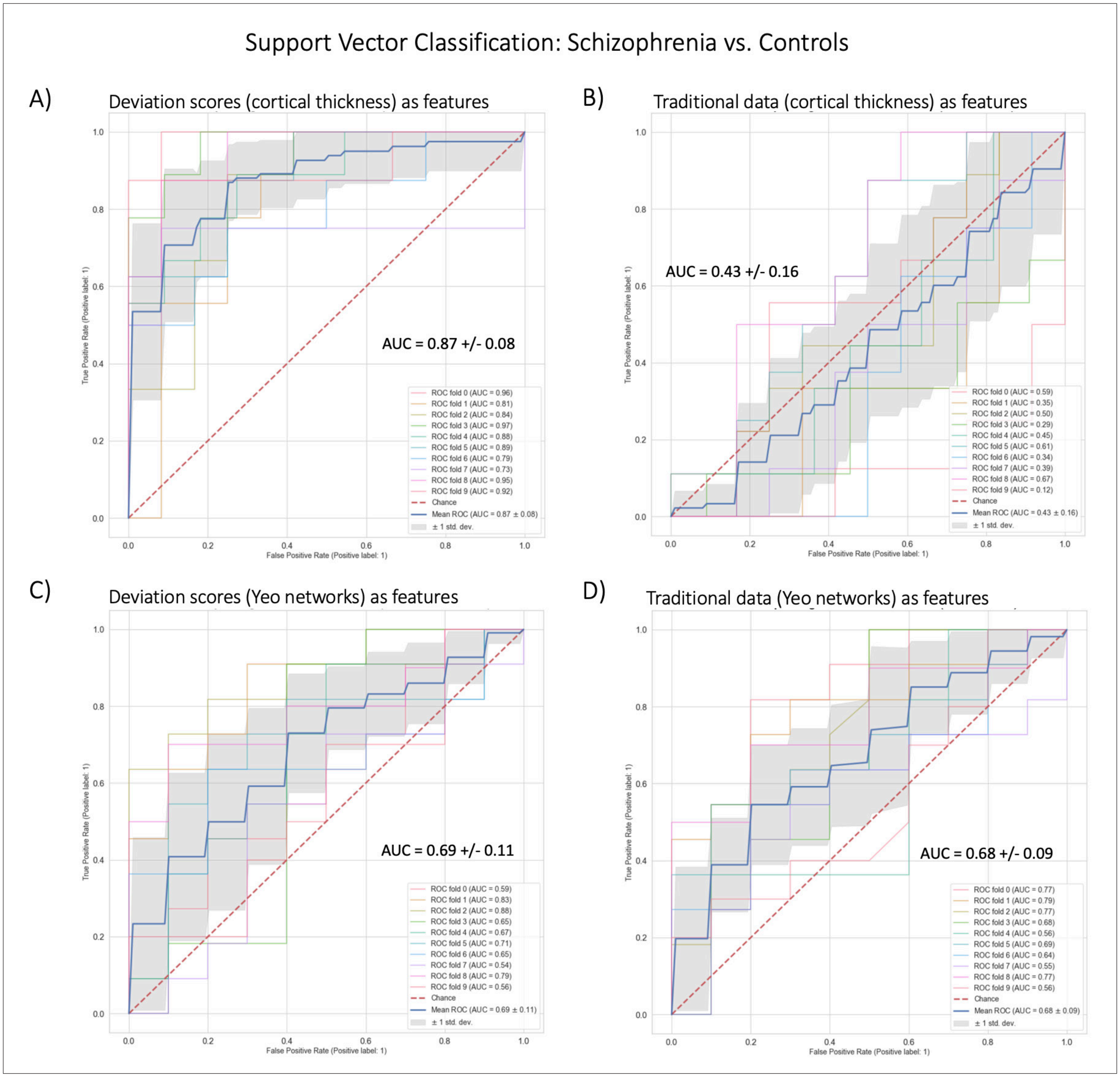


Figure 5. Benchmark task two multivariate prediction – Classification evaluation. (**A**) Support vector classification (SVC) using cortical thickness deviation scores as input features (most accurate model). (**B**) SVC using cortical thickness (residualized of sex and linear/quadratic effects of age) as input features. (**C**) SVC using functional brain network deviation scores as input features. (**D**) SVC using functional brain networks (residualized of sex and linear/quadratic effects of age and motion (mean framewise displacement)) as input features.

Normative modeling shows highest classification performance using cortical thickness

In benchmarking task two, we classified schizophrenia versus controls using SVC within a 10-fold cross-validation framework (***Table 2***, ***Figure 5***). The best-performing model used cortical thickness deviation scores to achieve a classification accuracy of 87% (AUC = 0.87). The raw cortical thickness model accuracy was indistinguishable from chance accuracy (AUC = 0.43). The AUC performance difference between the cortical thickness deviation and raw data models was 0.44, and this performance difference was statistically significant. The functional models, both deviation scores (0.69) and raw data (0.68) were more accurate than chance accuracy, however, the performance difference (i.e. improvement in accuracy using the deviation scores) was small (0.01) and was not statistically significant.

Normative modeling shows modest performance improvement in predicting cognition

In benchmarking task three we fit multivariate predictive models in a held-out test set of healthy individuals in the Human Connectome Project young-adult study to predict general cognitive ability (***Table 2***). The evidence provided by this task weakly favors the deviation score models. The most accurate (lowest mean squared error) model was the deviation cortical thickness model (MSE = 0.699). However, there was only an improvement of 0.008 in the deviation score model compared to the raw data model (MSE = 0.708) and this difference was not statistically significant. For the functional models, both the deviation score (MSE = 0.877) and raw data (MSE = 0.890) models were less accurate than the structural models and the difference between them (0.013) was also not statistically significant.

Discussion

This work expands the available open-source tools for conducting normative modeling analyses and provides clear evidence for why normative modeling should be utilized by the neuroimaging community (and beyond). We updated our publicly available repository of pre-trained normative models to include a new MRI imaging modality (models of resting-state functional connectivity extracted from the Yeo-17 and Smith-10 brain network atlases) and demonstrate how to transfer these models to new data sources. The repository includes an example transfer dataset and in addition, we have developed a user-friendly interface (https://pcnportal.dccn.nl/) that allows transferring the pre-trained normative models to new samples without requiring any programming. Next, we compared the features that are output from normative modeling (deviation scores) against 'raw' data features across several benchmarking tasks including univariate group difference testing (schizophrenia vs. control), multivariate prediction – classification (schizophrenia vs. control), and multivariate prediction – regression (predicting general cognitive ability). We found across all benchmarking tasks there were minor (regression) to strong (group difference testing) benefits of using deviation scores compared to the raw data features.

The fact that the deviation score models perform better than the raw data models confirms the utility of placing individuals into reference models. Our results show that normative modeling can capture population trends, uncover clinical group differences, and preserve the ability to study individual differences. We have some intuition on why the deviation score models perform better on the benchmarking tasks than the raw data. With normative modeling, we are accounting for many sources of variance that are not necessarily clinically meaningful (i.e. site) and we are able to capture clinically meaningful information within the reference cohort perspective. The reference model helps beyond just removing confounding variables such as scanner noise because we show that even when removing the nuisance covariates (age, sex, site, head motion) from the raw data, the normative modeling features still perform better on the benchmarking tasks.

Prior works on the methodological innovation and application of normative modeling ***Kia et al., 2018***; ***Kia et al., 2020***; ***Kia et al., 2021***; ***Kia and Marquand, 2018*** have focused on the beginning foundational steps of the framework (i.e. data selection and preparation, algorithmic implementation, and carefully evaluating out of sample model performance). However, the framework does not end after the model has been fit to the data (estimation step) and performance metrics have been established (evaluation step). Transferring the models to new samples, interpretation of the results, and potential downstream analysis are equally important steps, but they have received less attention. When it comes time to interpret the model outputs, it is easy to fall back into the case-control thinking paradigm, even after fitting a normative model to one's data (which is supposed to be an alternative to case vs. control approaches). This is due in part to the challenges arising from the results existing in a very high dimensional space (~100 s to 1000 s of brain regions from ~100 s to 1000 s of subjects). There is a reasonable need to distill and summarize these high-dimensional results. However, it is important to remember there is always a trade-off between having a complex enough of a model to explain the data and dimensionality reduction for the sake of interpretation simplicity. This distillation process often leads back to placing individuals into groups (i.e. case-control thinking) and interpreting group patterns or looking for group effects, rather than interpreting results at the level of the individual. We acknowledge the value and complementary nature of understanding individual variation relative to group means (case-control thinking) and clarify that we do not claim the superiority of normative modeling over case-control methods. Rather, our results from this work, especially in the comparisons of group difference map to individual difference maps (***Figure 4***), show that the outputs of normative modeling can be used to validate, refine, and further understand some of the inconsistencies in previous findings from case-control literature.

There are several limitations of the present work. First, the representation of functional normative models may be surprising and concerning. Typically, resting-state connectivity matrices are calculated using parcellations containing between 100–1000 nodes and 5000–500,000 connections. However, the Yeo-17 atlas (***Yeo et al., 2011***) was specifically chosen because of its widespread use and the fact that many other (higher resolution) functional brain parcellations have been mapped to the Yeo brain networks (***Eickhoff et al., 2018***; ***Glasser et al., 2016***; ***Gordon et al., 2016***; ***Gordon et al., 2017***; ***Kong et al., 2019***; ***Laumann et al., 2015***; ***Power et al., 2011***; ***Schaefer et al., 2018***; ***Shen et al., 2013***). There is an on-going debate about the best representation of functional brain activity. Using the Yeo-17 brain networks to model functional connectivity ignores important considerations regarding brain dynamics, flexible node configurations, overlapping functional modes, hard versus soft parcellations, and many other important issues. We have also shared functional normative models using the Smith-10 ICA-based parcellation, though did not repeat the benchmarking tasks using these data. Apart from our choice of parcellation, there are fundamental open questions regarding the nature of the brain's functional architecture, including how it is defined and measured. While it is outside the scope of this work to engage in these debates, we acknowledge their importance and refer curious readers to a thorough review of functional connectivity challenges (***Bijsterbosch et al., 2020***).

We would also like to expand on our prior discussion (***Rutherford et al., 2022a***) on the limitations of the reference cohort demographics, and the use of the word 'normative.' The included sample for training the functional normative models in this work, and the structural normative modeling sample in ***Rutherford et al., 2022a*** are most likely overrepresentative of European-ancestry (WEIRD population ***Henrich et al., 2010***) due to the data coming from academic research studies, which do not match global population demographics. Our models do not include race or ethnicity as covariates due to data availability (many sites did not provide race or ethnicity information). Prior research supports the use of age-specific templates and ethnicity-specific growth charts (***Dong et al., 2020***). This is a major limitation that requires additional future work and should be considered carefully when transferring the model to diverse data (***Benkarim et al., 2022***; ***Greene et al., 2022***; ***Li et al., 2022***). The term 'normative model' can be defined in other fields in a very different manner than ours (***Baron, 2004***; ***Colyvan, 2013***; ***Titelbaum, 2021***). We clarify that ours is strictly a statistical notion (normative = being within the central tendency for a population). Critically, we do not use normative in a moral or ethical sense, and we are not suggesting that individuals with high deviation scores require action or intervention to be pulled toward the population average. Although in some cases this may be true, we in no way assume that high deviations are problematic or unhealthy (they may in fact represent compensatory changes that are adaptive). In any case, we treat large deviations from statistical normality strictly as markers predictive of clinical states or conditions of interest.

There are many open research questions regarding normative modeling. Future research directions are likely to include: (1) further expansion of open-source pre-trained normative modeling repositories to include additional MRI imaging modalities such as task-based functional MRI and diffusion-weighted imaging, other neuroimaging modalities such as EEG or MEG, and models that include other non-biological measures, (2) increase in the resolution of existing models (i.e. voxel, vertex, models of brain structure and higher resolution functional parcellations), (3) replication and refinement of the proposed benchmarking tasks in other datasets including hyperparameter tuning and different algorithm implementation, and improving the regression benchmarking task, and (4) including additional benchmarking tasks beyond the ones considered here.

There has been recent interesting work on 'failure analysis' of brain-behavior models (***Greene et al., 2022***), and we would like to highlight that normative modeling is an ideal method for conducting this type of analysis. Through normative modeling, research questions such as 'what are the common patterns in the subjects that are classified well versus those that are not classified well' can be explored. Additional recent work (***Marek et al., 2022***) has highlighted important issues the brain-behavior modeling community must face, such as poor reliability of the imaging data, poor stability and accuracy of the predictive models, and the very large sample sizes (exceeding that of even the largest neuroimaging samples) required for accurate predictions. There has also been working showing that brain-behavior predictions are more reliable than the underlying functional data (***Taxali et al., 2021***), and other ideas for improving brain-behavior predictive models are discussed in-depth here (***Finn and Rosenberg, 2021***; ***Rosenberg and Finn, 2022***). Nevertheless, we acknowledge these challenges and believe that sharing pre-trained machine learning models and further development of transfer learning of these models could help further address these issues.

In this work, we have focused on the downstream steps of the normative modeling framework involving evaluation and interpretation, and how insights can be made on multiple levels. Through the precise modeling of different sources of variation, there is much knowledge to be gained at the level of populations, clinical groups, and individuals.

Additional information

Competing interests

Christian F Beckmann: is director and shareholder of SBGNeuro Ltd. Henricus G Ruhe: received speaker's honorarium from Lundbeck and Janssen. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
European Research Council	10100118	Andre F Marquand
Wellcome Trust	215698/Z/19/Z	Andre F Marquand
Wellcome Trust	098369/Z/12/Z	Christian F Beckmann
National Institute of Mental Health	R01MH122491	Ivy F Tso
National Institute of Mental Health	R01MH123458	Chandra Sripada
National Institute of Mental Health	R01MH130348	Chandra Sripada

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Author contributions

Saige Rutherford, Conceptualization, Resources, Data curation, Software, Formal analysis, Validation, Methodology, Writing - original draft, Writing – review and editing; Pieter Barkema, Resources, Software, Writing – review and editing; Ivy F Tso, Chandra Sripada, Christian F Beckmann, Henricus G Ruhe, Andre F Marquand, Supervision, Funding acquisition, Writing – review and editing

Author ORCIDs

Saige Rutherford http://orcid.org/0000-0003-3006-9044
Chandra Sripada http://orcid.org/0000-0001-9025-6453

Ethics

Human subjects: Secondary data analysis was conducted in this work. Data were pooled from multiple data sources described in the supplemental tables. All subjects provided informed consent. Subject recruitment procedures and informed consent forms, including consent to share de-identified data, were approved by the corresponding university institutional review board where data were collected. Human subjects: Ethical approval for the public data were provided by the relevant local research authorities for the studies contributing data. For full details, see the main study publications in the main text. For all clinical studies, approval was obtained via the local ethical review authorities, i.e., Delta: The local ethics committee of the Academic Medical Center of the University of Amsterdam (AMC-METC) Nr.:11/050, UMich_IMPS: University of Michigan Institution Review Board HUM00088188, UMich_SZG: University of Michigan Institution Review Board HUM00080457.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.85082.sa1
Author response https://doi.org/10.7554/eLife.85082.sa2

Additional files

Supplementary files

- Supplementary file 1. Functional Normative Model Demographics. *Description:* For each included site, we show the sample size (N), age (mean, standard deviation), and sex distribution (Female/Male percent) in the training set (shown in blue) and testing set (shown in green) of the normative models of functional connectivity between large scale resting-state brain networks from the Yeo 17 network atlas.
- Supplementary file 2. Surface Area Normative Model Demographics. *Description:* For each included site, we show the sample size (N), age (mean, standard deviation), and sex distribution (Female/Male percent) of the normative models of surface area extracted for all regions of interest in the Destrieux Freesurfer atlas.
- Supplementary file 3. Structural Group Difference Testing Statistics. *Description:* We show for all cortical thickness and subcortical volume from the Destrieux and aseg Freesurfer atlases regions of interest (ROIs from a two-sample t-test between Schizophrenia versus Healthy Controls) the t-statistic (T-stat), False Discovery Rate corrected p-value (FDRcorr_pvalue), and uncorrected p-value (uncorr_pvalue) for both the raw data (shown in green) and the deviation scores (shown in blue).
- Supplementary file 4. Functional Connectivity Group Difference Testing Statistics. *Description:* We show for all Yeo-17 between network connectivity regions of interest (ROIs) from a two-sample t-test between Schizophrenia versus Healthy Controls the t-statistic (T-stat), False Discovery Rate corrected p-value (FDRcorr_pvalue), and uncorrected p-value (uncorr_pvalue) for both the raw data (shown in green) and the deviation scores (shown in blue).
- MDAR checklist

Data availability

Pre-trained normative models are available on GitHub (https://github.com/predictive-clinical-neuroscience/braincharts, (copy archived at swh:1:rev:299e126ff053e2353091831a888c3ccd1ca6edeb)) and Google Colab (https://colab.research.google.com/github/predictive-clinical-neuroscience/braincharts/blob/master/scripts/apply_normative_models_yeo17.ipynb). Scripts for running the benchmarking analysis and visualizations are available on GitHub (https://github.com/saigerutherford/evidence_embracing_nm, (copy archived at swh:1:rev:1b4198389e2940dd3d10055164d68d46e0a20750)). An online portal for running models without code is available (https://pcnportal.dccn.nl).

References

Aine CJ, Bockholt HJ, Bustillo JR, Cañive JM, Caprihan A, Gasparovic C, Hanlon FM, Houck JM, Jung RE, Lauriello J, Liu J, Mayer AR, Perrone-Bizzozero NI, Posse S, Stephen JM, Turner JA, Clark VP, Calhoun VD.

2017. Multimodal neuroimaging in schizophrenia: description and dissemination. *Neuroinformatics* **15**:343–364. DOI: https://doi.org/10.1007/s12021-017-9338-9, PMID: 28812221

Arbabshirani MR, Kiehl KA, Pearlson GD, Calhoun VD. 2013. Classification of schizophrenia patients based on resting-state functional network connectivity. *Frontiers in Neuroscience* **7**:133. DOI: https://doi.org/10.3389/fnins.2013.00133, PMID: 23966903

Baker JT, Dillon DG, Patrick LM, Roffman JL, Brady RO, Pizzagalli DA, Öngür D, Holmes AJ. 2019. Functional connectomics of affective and psychotic pathology. *PNAS* **116**:9050–9059. DOI: https://doi.org/10.1073/pnas.1820780116, PMID: 30988201

Baron J. 2004. Normative models of judgment and decision making. Koehler DJ, Harvey N (Eds). *Blackwell Handbook of Judgment and Decision Making*. Blackwell Publishing Ltd. p. 19–36. DOI: https://doi.org/10.1002/9780470752937.ch2

Benkarim O, Paquola C, Park BY, Kebets V, Hong SJ, Vos de Wael R, Zhang S, Yeo BTT, Eickenberg M, Ge T, Poline JB, Bernhardt BC, Bzdok D. 2022. Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. *PLOS Biology* **20**:e3001627. DOI: https://doi.org/10.1371/journal.pbio.3001627, PMID: 35486643

Bijsterbosch J, Harrison SJ, Jbabdi S, Woolrich M, Beckmann C, Smith S, Duff EP. 2020. Challenges and future directions for representations of functional brain organization. *Nature Neuroscience* **23**:1484–1495. DOI: https://doi.org/10.1038/s41593-020-00726-z, PMID: 33106677

Borghi E, de Onis M, Garza C, Van den Broeck J, Frongillo EA, Grummer-Strawn L, Van Buuren S, Pan H, Molinari L, Martorell R, Onyango AW, Martines JC, WHO Multicentre Growth Reference Study Group. 2006. Construction of the world health organization child growth standards: selection of methods for attained growth curves. *Statistics in Medicine* **25**:247–265. DOI: https://doi.org/10.1002/sim.2227, PMID: 16143968

Cai N, Choi KW, Fried EI. 2020. Reviewing the genetics of heterogeneity in depression: operationalizations, manifestations and etiologies. *Human Molecular Genetics* **29**:R10–R18. DOI: https://doi.org/10.1093/hmg/ddaa115, PMID: 32568380

Castro E, Hjelm RD, Plis SM, Dinh L, Turner JA, Calhoun VD. 2016. Deep independence network analysis of structural brain imaging: application to schizophrenia. *IEEE Transactions on Medical Imaging* **35**:1729–1740. DOI: https://doi.org/10.1109/TMI.2016.2527717, PMID: 26891483

Cetin MS, Houck JM, Vergara VM, Miller RL, Calhoun V. 2015. Multimodal based classification of schizophrenia patients. *Conf Proc IEEE Eng Med Biol Soc* **2015**:2629–2632. DOI: https://doi.org/10.1109/EMBC.2015.7318931, PMID: 26736831

Cetin MS, Houck JM, Rashid B, Agacoglu O, Stephen JM, Sui J, Canive J, Mayer A, Aine C, Bustillo JR, Calhoun VD. 2016. Multimodal classification of schizophrenia patients with MEG and fmri data using static and dynamic connectivity measures. *Frontiers in Neuroscience* **10**:466. DOI: https://doi.org/10.3389/fnins.2016.00466, PMID: 27807403

Chang LJ, Gianaros PJ, Manuck SB, Krishnan A, Wager TD. 2015. A sensitive and specific neural signature for picture-induced negative affect. *PLOS Biology* **13**:e1002180. DOI: https://doi.org/10.1371/journal.pbio.1002180, PMID: 26098873

Cheon EJ, Bearden CE, Sun D, Ching CRK, Andreassen OA, Schmaal L, Veltman DJ, Thomopoulos SI, Kochunov P, Jahanshad N, Thompson PM, Turner JA, van Erp TGM. 2022. Cross disorder comparisons of brain structure in schizophrenia, bipolar disorder, major depressive disorder, and 22q11.2 deletion syndrome: a review of enigma findings. *Psychiatry and Clinical Neurosciences* **76**:140–161. DOI: https://doi.org/10.1111/pcn.13337, PMID: 35119167

Colyvan M. 2013. Idealisations in normative models. *Synthese* **190**:1337–1350. DOI: https://doi.org/10.1007/s11229-012-0166-z

Cuthbert BN, Insel TR. 2013. Toward the future of psychiatric diagnosis: the seven pillars of rdoc. *BMC Medicine* **11**:126. DOI: https://doi.org/10.1186/1741-7015-11-126, PMID: 23672542

Dale AM, Fischl B, Sereno MI. 1999. Cortical surface-based analysis. I. segmentation and surface reconstruction. *NeuroImage* **9**:179–194. DOI: https://doi.org/10.1006/nimg.1998.0395, PMID: 9931268

Dansereau C, Benhajali Y, Risterucci C, Pich EM, Orban P, Arnold D, Bellec P. 2017. Statistical power and prediction accuracy in multisite resting-state fmri connectivity. *NeuroImage* **149**:220–232. DOI: https://doi.org/10.1016/j.neuroimage.2017.01.072, PMID: 28161310

Di Biase MA, Cropley VL, Cocchi L, Fornito A, Calamante F, Ganella EP, Pantelis C, Zalesky A. 2019. Linking cortical and connectional pathology in schizophrenia. *Schizophrenia Bulletin* **45**:911–923. DOI: https://doi.org/10.1093/schbul/sby121, PMID: 30215783

Dinga R, Fraza CJ, Bayer JMM, Kia SM, Beckmann CF, Marquand AF. 2021. Normative Modeling of Neuroimaging Data Using Generalized Additive Models of Location Scale and Shape. *bioRxiv*. DOI: https://doi.org/10.1101/2021.06.14.448106

Dong H-M, Castellanos FX, Yang N, Zhang Z, Zhou Q, He Y, Zhang L, Xu T, Holmes AJ, Thomas Yeo BT, Chen F, Wang B, Beckmann C, White T, Sporns O, Qiu J, Feng T, Chen A, Liu X, Chen X, et al. 2020. Charting brain growth in tandem with brain templates at school age. *Science Bulletin* **65**:1924–1934. DOI: https://doi.org/10.1016/j.scib.2020.07.027, PMID: 36738058

Dwyer DB, Cabral C, Kambeitz-Ilankovic L, Sanfelici R, Kambeitz J, Calhoun V, Falkai P, Pantelis C, Meisenzahl E, Koutsouleris N. 2018. Brain subtyping enhances the neuroanatomical discrimination of schizophrenia. *Schizophrenia Bulletin* **44**:1060–1069. DOI: https://doi.org/10.1093/schbul/sby008, PMID: 29529270

Eickhoff SB, Yeo BTT, Genon S. 2018. Imaging-based parcellations of the human brain. *Nature Reviews. Neuroscience* **19**:672–686. DOI: https://doi.org/10.1038/s41583-018-0071-7, PMID: 30305712

Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, Oya H, Ghosh SS, Wright J, Durnez J, Poldrack RA, Gorgolewski KJ. 2019. FMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods* **16**:111–116. DOI: https://doi.org/10.1038/s41592-018-0235-4, PMID: 30532080

Etkin A. 2019. A reckoning and research agenda for neuroimaging in psychiatry. *The American Journal of Psychiatry* **176**:507–511. DOI: https://doi.org/10.1176/appi.ajp.2019.19050521, PMID: 31256624

Filip P, Bednarik P, Eberly LE, Moheet A, Svatkova A, Grohn H, Kumar AF, Seaquist ER, Mangia S. 2022. Different freesurfer versions might generate different statistical outcomes in case-control comparison studies. *Neuroradiology* **64**:765–773. DOI: https://doi.org/10.1007/s00234-021-02862-0, PMID: 34988592

Finn ES, Rosenberg MD. 2021. Beyond fingerprinting: choosing predictive connectomes over reliable connectomes. *NeuroImage* **239**:118254. DOI: https://doi.org/10.1016/j.neuroimage.2021.118254, PMID: 34118397

First MB. 1956. In SCID-5-CV: Structured Clinical Interview for DSM-5 Disorders: Clinician Version. American Psychiatric Association Publishing; U-M Catalog Search.

Fischl B, Dale AM. 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *PNAS* **97**:11050–11055. DOI: https://doi.org/10.1073/pnas.200033797, PMID: 10984517

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM. 2002. Whole brain segmentation. *Neuron* **33**:341–355. DOI: https://doi.org/10.1016/S0896-6273(02)00569-X, PMID: 11832223

Flake JK, Fried EI. 2020. Measurement schmeasurement: questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science* **3**:456–465. DOI: https://doi.org/10.1177/2515245920952393

Floris DL, Wolfers T, Zabihi M, Holz NE, Zwiers MP, Charman T, Tillmann J, Ecker C, Dell'Acqua F, Banaschewski T, Moessnang C, Baron-Cohen S, Holt R, Durston S, Loth E, Murphy DGM, Marquand A, Buitelaar JK, Beckmann CF, EU-AIMS Longitudinal European Autism Project Group. 2021. Atypical brain asymmetry in autism-A candidate for clinically meaningful stratification. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging* **6**:802–812. DOI: https://doi.org/10.1016/j.bpsc.2020.08.008, PMID: 33097470

Fraza CJ, Dinga R, Beckmann CF, Marquand AF. 2021. Warped bayesian linear regression for normative modelling of big data. *NeuroImage* **245**:118715. DOI: https://doi.org/10.1016/j.neuroimage.2021.118715, PMID: 34798518

Fraza C, Zabihi M, Beckmann CF, Marquand AF. 2022. The Extremes of Normative Modelling. *bioRxiv*. DOI: https://doi.org/10.1101/2022.08.23.505049

Gau R, Noble S, Heuer K, Bottenhorn KL, Bilgin IP, Yang Y-F, Huntenburg JM, Bayer JMM, Bethlehem RAI, Rhoads SA, Vogelbacher C, Borghesani V, Levitis E, Wang H-T, Van Den Bossche S, Kobeleva X, Legarreta JH, Guay S, Atay SM, Varoquaux GP, et al. 2021. Brainhack: developing a culture of open, inclusive, community-driven neuroscience. *Neuron* **109**:1769–1775. DOI: https://doi.org/10.1016/j.neuron.2021.04.001, PMID: 33932337

Geisler D, Walton E, Naylor M, Roessner V, Lim KO, Charles Schulz S, Gollub RL, Calhoun VD, Sponheim SR, Ehrlich S. 2015. Brain structure and function correlates of cognitive subtypes in schizophrenia. *Psychiatry Research* **234**:74–83. DOI: https://doi.org/10.1016/j.pscychresns.2015.08.008, PMID: 26341950

Gershon RC, Cella D, Fox NA, Havlik RJ, Hendrie HC, Wagster MV. 2010. Assessment of neurological and behavioural function: the NIH toolbox. *The Lancet. Neurology* **9**:138–139. DOI: https://doi.org/10.1016/S1474-4422(09)70335-7, PMID: 20129161

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, Van Essen DC. 2016. A multi-modal parcellation of human cerebral cortex. *Nature* **536**:171–178. DOI: https://doi.org/10.1038/nature18933, PMID: 27437579

Goodkind M, Eickhoff SB, Oathes DJ, Jiang Y, Chang A, Jones-Hagata LB, Ortega BN, Zaiko YV, Roach EL, Korgaonkar MS, Grieve SM, Galatzer-Levy I, Fox PT, Etkin A. 2015. Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry* **72**:305–315. DOI: https://doi.org/10.1001/jamapsychiatry.2014.2206, PMID: 25651064

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. *Cerebral Cortex* **26**:288–303. DOI: https://doi.org/10.1093/cercor/bhu239, PMID: 25316338

Gordon EM, Laumann TO, Gilmore AW, Newbold DJ, Greene DJ, Berg JJ, Ortega M, Hoyt-Drazen C, Gratton C, Sun H, Hampton JM, Coalson RS, Nguyen AL, McDermott KB, Shimony JS, Snyder AZ, Schlaggar BL, Petersen SE, Nelson SM, Dosenbach NUF. 2017. Precision functional mapping of individual human brains. *Neuron* **95**:791–807. DOI: https://doi.org/10.1016/j.neuron.2017.07.011, PMID: 28757305

Greene AS, Shen X, Noble S, Horien C, Hahn CA, Arora J, Tokoglu F, Spann MN, Carrión CI, Barron DS, Sanacora G, Srihari VH, Woods SW, Scheinost D, Constable RT. 2022. Brain-phenotype models fail for individuals who defy sample stereotypes. *Nature* **609**:109–118. DOI: https://doi.org/10.1038/s41586-022-05118-w, PMID: 36002572

Henrich J, Heine SJ, Norenzayan A. 2010. The weirdest people in the world? *The Behavioral and Brain Sciences* **33**:61–83. DOI: https://doi.org/10.1017/S0140525X0999152X, PMID: 20550733

Holz NE, Floris DL, Llera A, Aggensteiner PM, Kia SM, Wolfers T, Baumeister S, Böttinger B, Glennon JC, Hoekstra PJ, Dietrich A, Saam MC, Schulze UME, Lythgoe DJ, Williams SCR, Santosh P, Rosa-Justicia M, Bargallo N, Castro-Fornieles J, Arango C, et al. 2022. Age-related brain deviations and aggression. *Psychological Medicine* **1**:1–10. DOI: https://doi.org/10.1017/S003329172200068X, PMID: 35450543

Howes OD, Cummings C, Chapman GE, Shatalina E. 2023. Neuroimaging in schizophrenia: an overview of findings and their implications for synaptic changes. *Neuropsychopharmacology* **48**:151–167. DOI: https://doi.org/10.1038/s41386-022-01426-x, PMID: 36056106

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, Sanislow C, Wang P. 2010. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry* **167**:748–751. DOI: https://doi.org/10.1176/appi.ajp.2010.09091379, PMID: 20595427

Itälinna V, Kaltiainen H, Forss N, Liljeström M, Parkkonen L. 2022. Detecting Mild Traumatic Brain Injury with MEG, Normative Modelling and Machine Learning. *medRxiv*. DOI: https://doi.org/10.1101/2022.09.29.22280521

Jolliffe IT. 1982. A note on the use of principal components in regression. *Applied Statistics* **31**:300. DOI: https://doi.org/10.2307/2348005

Jones MC, Pewsey A. 2009. Sinh-arcsinh distributions. *Biometrika* **96**:761–780. DOI: https://doi.org/10.1093/biomet/asp053

Kia SM, Beckmann CF, Marquand AF. 2018. Scalable Multi-Task Gaussian Process Tensor Regression for Normative Modeling of Structured Variation in Neuroimaging Data. *arXiv*. http://arxiv.org/abs/1808.00036

Kia SM, Marquand A. 2018. Normative Modeling of Neuroimaging Data Using Scalable Multi-Task Gaussian Processes. *arXiv*. http://arxiv.org/abs/1806.01047

Kia SM, Huijsdens H, Dinga R, Wolfers T, Mennes M, Andreassen OA, Westlye LT, Beckmann CF, Marquand AF. 2020. Hierarchical bayesian regression for multi-site normative modeling of neuroimaging data. Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L (Eds). *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Springer International Publishing. p. 699–709. DOI: https://doi.org/10.1007/978-3-030-59728-3_68

Kia SM, Huijsdens H, Rutherford S, Dinga R, Wolfers T, Mennes M, Andreassen OA, Westlye LT, Beckmann CF, Marquand AF. 2021. Federated Multi-Site Normative Modeling Using Hierarchical Bayesian Regression. *bioRxiv*. DOI: https://doi.org/10.1101/2021.05.28.446120

Kia SM, Huijsdens H, Rutherford S, de Boer A, Dinga R, Wolfers T, Berthet P, Mennes M, Andreassen OA, Westlye LT, Beckmann CF, Marquand AF. 2022. Closing the life-cycle of normative modeling using federated hierarchical bayesian regression. *PLOS ONE* **17**:e0278776. DOI: https://doi.org/10.1371/journal.pone.0278776, PMID: 36480551

Kjelkenes R, Wolfers T, Alnæs D, van der Meer D, Pedersen ML, Dahl A, Voldsbekk I, Moberget T, Tamnes CK, Andreassen OA, Marquand AF, Westlye LT. 2022. Mapping normative trajectories of cognitive function and its relation to psychopathology symptoms and genetic risk in youth. *Biological Psychiatry Global Open Science* **1**:007. DOI: https://doi.org/10.1016/j.bpsgos.2022.01.007

Klapwijk ET, van de Kamp F, van der Meulen M, Peters S, Wierenga LM. 2019. Qoala-T: a supervised-learning tool for quality control of freesurfer segmented MRI data. *NeuroImage* **189**:116–129. DOI: https://doi.org/10.1016/j.neuroimage.2019.01.014, PMID: 30633965

Kong R, Li J, Orban C, Sabuncu MR, Liu H, Schaefer A, Sun N, Zuo XN, Holmes AJ, Eickhoff SB, Yeo BTT. 2019. Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cerebral Cortex* **29**:2533–2551. DOI: https://doi.org/10.1093/cercor/bhy123, PMID: 29878084

Kumar S. 2021. NormVAE: Normative Modeling on Neuroimaging Data Using Variational Autoencoders. *arXiv*. https://arxiv.org/abs/2110.04903v2

Laumann TO, Gordon EM, Adeyemo B, Snyder AZ, Joo SJ, Chen MY, Gilmore AW, McDermott KB, Nelson SM, Dosenbach NUF, Schlaggar BL, Mumford JA, Poldrack RA, Petersen SE. 2015. Functional system and areal organization of a highly sampled individual human brain. *Neuron* **87**:657–670. DOI: https://doi.org/10.1016/j.neuron.2015.06.037, PMID: 26212711

Lee W, Bindman J, Ford T, Glozier N, Moran P, Stewart R, Hotopf M. 2007. Bias in psychiatric case-control studies: literature survey. *The British Journal of Psychiatry* **190**:204–209. DOI: https://doi.org/10.1192/bjp.bp.106.027250, PMID: 17329739

Lei D, Pinaya WHL, van Amelsvoort T, Marcelis M, Donohoe G, Mothersill DO, Corvin A, Gill M, Vieira S, Huang X, Lui S, Scarpazza C, Young J, Arango C, Bullmore E, Qiyong G, McGuire P, Mechelli A. 2020a. Detecting schizophrenia at the level of the individual: relative diagnostic value of whole-brain images, connectome-wide functional connectivity and graph-based metrics. *Psychological Medicine* **50**:1852–1861. DOI: https://doi.org/10.1017/S0033291719001934, PMID: 31391132

Lei D, Pinaya WHL, Young J, van Amelsvoort T, Marcelis M, Donohoe G, Mothersill DO, Corvin A, Vieira S, Huang X, Lui S, Scarpazza C, Arango C, Bullmore E, Gong Q, McGuire P, Mechelli A. 2020b. Integrating machining learning and multimodal neuroimaging to detect schizophrenia at the level of the individual. *Human Brain Mapping* **41**:1119–1135. DOI: https://doi.org/10.1002/hbm.24863, PMID: 31737978

Levitis E, van Praag CDG, Gau R, Heunis S, DuPre E, Kiar G, Bottenhorn KL, Glatard T, Nikolaidis A, Whitaker KJ, Mancini M, Niso G, Afyouni S, Alonso-Ortiz E, Appelhoff S, Arnatkeviciute A, Atay SM, Auer T, Baracchini G, Bayer JMM, et al. 2021. Centering inclusivity in the design of online conferences-an OHBM-open science perspective. *GigaScience* **10**:giab051. DOI: https://doi.org/10.1093/gigascience/giab051, PMID: 34414422

Li J, Bzdok D, Chen J, Tam A, Ooi LQR, Holmes AJ, Ge T, Patil KR, Jabbi M, Eickhoff SB, Yeo BTT, Genon S. 2022. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances* **8**:eabj1812. DOI: https://doi.org/10.1126/sciadv.abj1812, PMID: 35294251

Linden DEJ. 2012. The challenges and promise of neuroimaging in psychiatry. *Neuron* **73**:8–22. DOI: https://doi.org/10.1016/j.neuron.2011.12.014, PMID: 22243743

Loth E, Ahmad J, Chatham C, López B, Carter B, Crawley D, Oakley B, Hayward H, Cooke J, San José Cáceres A, Bzdok D, Jones E, Charman T, Beckmann C, Bourgeron T, Toro R, Buitelaar J, Murphy D, Dumas G. 2021. The meaning of significant mean group differences for biomarker discovery. *PLOS Computational Biology* **17**:e1009477. DOI: https://doi.org/10.1371/journal.pcbi.1009477, PMID: 34793435

Lv J, Di Biase M, Cash RFH, Cocchi L, Cropley VL, Klauser P, Tian Y, Bayer J, Schmaal L, Cetin-Karayumak S, Rathi Y, Pasternak O, Bousman C, Pantelis C, Calamante F, Zalesky A. 2021. Individual deviations from normative models of brain structure in a large cross-sectional schizophrenia cohort. *Molecular Psychiatry* **26**:3512–3523. DOI: https://doi.org/10.1038/s41380-020-00882-5, PMID: 32963336

Madre M, Canales-Rodríguez EJ, Fuentes-Claramonte P, Alonso-Lana S, Salgado-Pineda P, Guerrero-Pedraza A, Moro N, Bosque C, Gomar JJ, Ortíz-Gil J, Goikolea JM, Bonnin CM, Vieta E, Sarró S, Maristany T, McKenna PJ, Salvador R, Pomarol-Clotet E. 2020. Structural abnormality in schizophrenia versus bipolar disorder: a whole brain cortical thickness, surface area, volume and gyrification analyses. *NeuroImage. Clinical* **25**:102131. DOI: https://doi.org/10.1016/j.nicl.2019.102131, PMID: 31911343

Mann HB, Whitney DR. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* **18**:50–60. DOI: https://doi.org/10.1214/aoms/1177730491

Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Donohue MR, Foran W, Miller RL, Hendrickson TJ, Malone SM, Kandala S, Feczko E, Miranda-Dominguez O, Graham AM, Earl EA, Perrone AJ, Cordova M, Doyle O, Moore LA, et al. 2022. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**:654–660. DOI: https://doi.org/10.1038/s41586-022-04492-9, PMID: 35296861

Marquand AF, Rezek I, Buitelaar J, Beckmann CF. 2016a. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological Psychiatry* **80**:552–561. DOI: https://doi.org/10.1016/j.biopsych.2015.12.023, PMID: 26927419

Marquand AF, Wolfers T, Mennes M, Buitelaar J, Beckmann CF. 2016b. Beyond lumping and splitting: A review of computational approaches for stratifying psychiatric disorders. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging* **1**:433–447. DOI: https://doi.org/10.1016/j.bpsc.2016.04.002, PMID: 27642641

Marquand AF, Haak KV, Beckmann CF. 2017. Functional corticostriatal connection topographies predict goal directed behaviour in humans. *Nature Human Behaviour* **1**:0146. DOI: https://doi.org/10.1038/s41562-017-0146, PMID: 28804783

Marquand A, Rutherford S, Kia SM, Wolfers T, Fraza C, Dinga R, Zabihi M. 2021. PCNToolkit. Zenodo. https://doi.org/10.5281/zenodo.5207839

McTeague LM, Huemer J, Carreon DM, Jiang Y, Eickhoff SB, Etkin A. 2017. Identification of common neural circuit disruptions in cognitive control across psychiatric disorders. *The American Journal of Psychiatry* **174**:676–685. DOI: https://doi.org/10.1176/appi.ajp.2017.16040400, PMID: 28320224

Meng X, Jiang R, Lin D, Bustillo J, Jones T, Chen J, Yu Q, Du Y, Zhang Y, Jiang T, Sui J, Calhoun VD. 2017. Predicting individualized clinical measures by a generalized prediction framework and multimodal fusion of MRI data. *NeuroImage* **145**:218–229. DOI: https://doi.org/10.1016/j.neuroimage.2016.05.026, PMID: 27177764

Michelini G, Palumbo IM, DeYoung CG, Latzman RD, Kotov R. 2021. Linking rdoc and hitop: a new interface for advancing psychiatric nosology and neuroscience. *Clinical Psychology Review* **86**:102025. DOI: https://doi.org/10.1016/j.cpr.2021.102025, PMID: 33798996

Moriarity DP, Alloy LB. 2021. Back to basics: the importance of measurement properties in biological psychiatry. *Neuroscience and Biobehavioral Reviews* **123**:72–82. DOI: https://doi.org/10.1016/j.neubiorev.2021.01.008, PMID: 33497789

Moriarity DP, Joyner KJ, Slavich GM, Alloy LB. 2022. Unconsidered issues of measurement noninvariance in biological psychiatry: a focus on biological phenotypes of psychopathology. *Molecular Psychiatry* **27**:1281–1285. DOI: https://doi.org/10.1038/s41380-021-01414-5, PMID: 34997192

Mottron L, Bzdok D. 2022. Diagnosing as autistic people increasingly distant from prototypes lead neither to clinical benefit nor to the advancement of knowledge. *Molecular Psychiatry* **27**:773–775. DOI: https://doi.org/10.1038/s41380-021-01343-3, PMID: 34642453

Nour MM, Liu Y, Dolan RJ. 2022. Functional neuroimaging in psychiatry and the case for failing better. *Neuron* **110**:2524–2544. DOI: https://doi.org/10.1016/j.neuron.2022.07.005, PMID: 35981525

Park SH. 1981. Collinearity and optimal restrictions on regression parameters for estimating responses. *Technometrics* **23**:289. DOI: https://doi.org/10.2307/1267793

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12**:2825–2830.

Pereira-Sanchez V, Castellanos FX. 2021. Neuroimaging in attention-deficit/hyperactivity disorder. *Current Opinion in Psychiatry* **34**:105–111. DOI: https://doi.org/10.1097/YCO.0000000000000669, PMID: 33278156

Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, Vogel AC, Laumann TO, Miezin FM, Schlaggar BL, Petersen SE. 2011. Functional network organization of the human brain. *Neuron* **72**:665–678. DOI: https://doi.org/10.1016/j.neuron.2011.09.006, PMID: 22099467

Pruim RHR, Mennes M, Buitelaar JK, Beckmann CF. 2015a. Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fmri. *NeuroImage* **112**:278–287. DOI: https://doi.org/10.1016/j.neuroimage.2015.02.063, PMID: 25770990

Pruim RHR, Mennes M, van Rooij D, Llera A, Buitelaar JK, Beckmann CF. 2015b. ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fmri data. *NeuroImage* **112**:267–277. DOI: https://doi.org/10.1016/j.neuroimage.2015.02.064, PMID: 25770991

Rahim M, Thirion B, Bzdok D, Buvat I, Varoquaux G. 2017. Joint prediction of multiple scores captures better individual traits from brain images. *NeuroImage* **158**:145–154. DOI: https://doi.org/10.1016/j.neuroimage.2017.06.072, PMID: 28676298

Rosa MJ, Portugal L, Hahn T, Fallgatter AJ, Garrido MI, Shawe-Taylor J, Mourao-Miranda J. 2015. Sparse network-based models for patient classification using fmri. *NeuroImage* **105**:493–506. DOI: https://doi.org/10.1016/j.neuroimage.2014.11.021, PMID: 25463459

Rosenberg MD, Finn ES. 2022. How to establish robust brain-behavior relationships without thousands of individuals. *Nature Neuroscience* **25**:835–837. DOI: https://doi.org/10.1038/s41593-022-01110-9, PMID: 35710985

Rutherford S, Angstadt M, Sripada C, Chang SE. 2020. Leveraging Big Data for Classification of Children Who Stutter from Fluent Peers. *bioRxiv*. DOI: https://doi.org/10.1101/2020.10.28.359711

Rutherford S, Fraza C, Dinga R, Kia SM, Wolfers T, Zabihi M, Berthet P, Worker A, Verdi S, Andrews D, Han LK, Bayer JM, Dazzan P, McGuire P, Mocking RT, Schene A, Sripada C, Tso IF, Duval ER, Chang SE, et al. 2022a. Charting brain growth and aging at high spatial precision. *eLife* **11**:e72904. DOI: https://doi.org/10.7554/eLife.72904, PMID: 35101172

Rutherford S, Kia SM, Wolfers T, Fraza C, Zabihi M, Dinga R, Berthet P, Worker A, Verdi S, Ruhe HG, Beckmann CF, Marquand AF. 2022b. The normative modeling framework for computational psychiatry. *Nature Protocols* **17**:1711–1734. DOI: https://doi.org/10.1038/s41596-022-00696-5, PMID: 35650452

Salvador R, Radua J, Canales-Rodríguez EJ, Solanes A, Sarró S, Goikolea JM, Valiente A, Monté GC, Natividad MDC, Guerrero-Pedraza A, Moro N, Fernández-Corcuera P, Amann BL, Maristany T, Vieta E, McKenna PJ, Pomarol-Clotet E. 2017. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLOS ONE* **12**:e0175683. DOI: https://doi.org/10.1371/journal.pone.0175683, PMID: 28426817

Sanislow CA. 2020. RDoC at 10: changing the discourse for psychopathology. *World Psychiatry* **19**:311–312. DOI: https://doi.org/10.1002/wps.20800, PMID: 32931117

Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, Eickhoff SB, Yeo BTT. 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex* **28**:3095–3114. DOI: https://doi.org/10.1093/cercor/bhx179, PMID: 28981612

Shen X, Tokoglu F, Papademetris X, Constable RT. 2013. Groupwise whole-brain parcellation from resting-state fmri data for network node identification. *NeuroImage* **82**:403–415. DOI: https://doi.org/10.1016/j.neuroimage.2013.05.081, PMID: 23747961

Shi D, Li Y, Zhang H, Yao X, Wang S, Wang G, Ren K. 2021. Machine learning of schizophrenia detection with structural and functional neuroimaging. *Disease Markers* **2021**:9963824. DOI: https://doi.org/10.1155/2021/9963824, PMID: 34211615

Siegel JS, Mitra A, Laumann TO, Seitzman BA, Raichle M, Corbetta M, Snyder AZ. 2017. Data quality influences observed links between functional connectivity and behavior. *Cerebral Cortex* **27**:4492–4502. DOI: https://doi.org/10.1093/cercor/bhw253, PMID: 27550863

Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, Filippini N, Watkins KE, Toro R, Laird AR, Beckmann CF. 2009. Correspondence of the brain's functional architecture during activation and rest. *PNAS* **106**:13040–13045. DOI: https://doi.org/10.1073/pnas.0905267106, PMID: 19620724

Sprooten E, Rasgon A, Goodman M, Carlin A, Leibu E, Lee WH, Frangou S. 2017. Addressing reverse inference in psychiatric neuroimaging: meta-analyses of task-related brain activation in common mental disorders. *Human Brain Mapping* **38**:1846–1864. DOI: https://doi.org/10.1002/hbm.23486, PMID: 28067006

Sripada C, Angstadt M, Rutherford S, Kessler D, Kim Y, Yee M, Levina E. 2019. Basic units of inter-individual variation in resting state connectomes. *Scientific Reports* **9**:1900. DOI: https://doi.org/10.1038/s41598-018-38406-5, PMID: 30760808

Sripada C, Angstadt M, Rutherford S, Taxali A, Shedden K. 2020a. Toward a "treadmill test" for cognition: improved prediction of general cognitive ability from the task activated brain. *Human Brain Mapping* **41**:3186–3197. DOI: https://doi.org/10.1002/hbm.25007, PMID: 32364670

Sripada C, Rutherford S, Angstadt M, Thompson WK, Luciana M, Weigard A, Hyde LH, Heitzeg M. 2020b. Prediction of neurocognition in youth from resting state fmri. *Molecular Psychiatry* **25**:3413–3421. DOI: https://doi.org/10.1038/s41380-019-0481-6, PMID: 31427753

Sui J, Pearlson GD, Du Y, Yu Q, Jones TR, Chen J, Jiang T, Bustillo J, Calhoun VD. 2015. In search of multimodal neuroimaging biomarkers of cognitive deficits in schizophrenia. *Biological Psychiatry* **78**:794–804. DOI: https://doi.org/10.1016/j.biopsych.2015.02.017, PMID: 25847180

Sui J, Qi S, van Erp TGM, Bustillo J, Jiang R, Lin D, Turner JA, Damaraju E, Mayer AR, Cui Y, Fu Z, Du Y, Chen J, Potkin SG, Preda A, Mathalon DH, Ford JM, Voyvodic J, Mueller BA, Belger A, et al. 2018. Multimodal neuromarkers in schizophrenia via cognition-guided MRI fusion. *Nature Communications* **9**:3028. DOI: https://doi.org/10.1038/s41467-018-05432-w, PMID: 30072715

Taxali A, Angstadt M, Rutherford S, Sripada C. 2021. Boost in test-retest reliability in resting state fmri with predictive modeling. *Cerebral Cortex* **31**:2822–2833. DOI: https://doi.org/10.1093/cercor/bhaa390, PMID: 33447841

Titelbaum MG. 2021. Normative Modeling [Preprint]. http://philsci-archive.pitt.edu/18670/ [Accessed February 1, 2021].

Tso IF, Angstadt M, Rutherford S, Peltier S, Diwadkar VA, Taylor SF. 2021. Dynamic causal modeling of eye gaze processing in schizophrenia. *Schizophrenia Research* **229**:112–121. DOI: https://doi.org/10.1016/j.schres.2020.11.012, PMID: 33229223

van Erp TGM, Walton E, Hibar DP, Schmaal L, Jiang W, Glahn DC, Pearlson GD, Yao N, Fukunaga M, Hashimoto R, Okada N, Yamamori H, Bustillo JR, Clark VP, Agartz I, Mueller BA, Cahn W, de Zwarte SMC, Hulshoff Pol HE, Kahn RS, et al. 2018. Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through meta analysis (enigma) consortium. *Biological Psychiatry* **84**:644–654. DOI: https://doi.org/10.1016/j.biopsych.2018.04.023, PMID: 29960671

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, Consortium WMH. 2013. The WU-minn human connectome project: an overview. *NeuroImage* **80**:62–79. DOI: https://doi.org/10.1016/j.neuroimage.2013.05.041, PMID: 23684880

van Haren NEM, Schnack HG, Cahn W, van den Heuvel MP, Lepage C, Collins L, Evans AC, Hulshoff Pol HE, Kahn RS. 2011. Changes in cortical thickness during the course of illness in schizophrenia. *Archives of General Psychiatry* **68**:871–880. DOI: https://doi.org/10.1001/archgenpsychiatry.2011.88, PMID: 21893656

Venkataraman A, Whitford TJ, Westin CF, Golland P, Kubicki M. 2012. Whole brain resting state functional connectivity abnormalities in schizophrenia. *Schizophrenia Research* **139**:7–12. DOI: https://doi.org/10.1016/j.schres.2012.04.021, PMID: 22633528

Verdi S, Marquand AF, Schott JM, Cole JH. 2021. Beyond the average patient: how neuroimaging models can address heterogeneity in dementia. *Brain* **144**:2946–2953. DOI: https://doi.org/10.1093/brain/awab165, PMID: 33892488

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods* **17**:261–272. DOI: https://doi.org/10.1038/s41592-019-0686-2, PMID: 32015543

Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E. 2013. An fmri-based neurologic signature of physical pain. *The New England Journal of Medicine* **368**:1388–1397. DOI: https://doi.org/10.1056/NEJMoa1204471, PMID: 23574118

Wannan CMJ, Cropley VL, Chakravarty MM, Bousman C, Ganella EP, Bruggemann JM, Weickert TW, Weickert CS, Everall I, McGorry P, Velakoulis D, Wood SJ, Bartholomeusz CF, Pantelis C, Zalesky A. 2019. Evidence for network-based cortical thickness reductions in schizophrenia. *The American Journal of Psychiatry* **176**:552–563. DOI: https://doi.org/10.1176/appi.ajp.2019.18040380, PMID: 31164006

Winter NR, Leenings R, Ernsting J, Sarink K, Fisch L, Emden D, Blanke J, Goltermann J, Opel N, Barkhau C, Meinert S, Dohm K, Repple J, Mauritz M, Gruber M, Leehr EJ, Grotegerd D, Redlich R, Jansen A, Nenadic I, et al. 2022. Quantifying deviations of brain structure and function in major depressive disorder across neuroimaging modalities. *JAMA Psychiatry* **79**:879–888. DOI: https://doi.org/10.1001/jamapsychiatry.2022.1780, PMID: 35895072

Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Biobehavioral Reviews* **57**:328–349. DOI: https://doi.org/10.1016/j.neubiorev.2015.08.001, PMID: 26254595

Wolfers T, Arenas AL, Onnink AMH, Dammers J, Hoogman M, Zwiers MP, Buitelaar JK, Franke B, Marquand AF, Beckmann CF. 2017. Refinement by integration: aggregated effects of multimodal imaging markers on adult ADHD. *Journal of Psychiatry & Neuroscience* **42**:386–394. DOI: https://doi.org/10.1503/jpn.160240, PMID: 28832320

Wolfers T, Doan NT, Kaufmann T, Alnæs D, Moberget T, Agartz I, Buitelaar JK, Ueland T, Melle I, Franke B, Andreassen OA, Beckmann CF, Westlye LT, Marquand AF. 2018. Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA Psychiatry* **75**:1146–1155. DOI: https://doi.org/10.1001/jamapsychiatry.2018.2467, PMID: 30304337

Wolfers T, Beckmann CF, Hoogman M, Buitelaar JK, Franke B, Marquand AF. 2020. Individual differences *v.* the average patient: mapping the heterogeneity in ADHD using normative models. *Psychological Medicine* **50**:314–323. DOI: https://doi.org/10.1017/S0033291719000084, PMID: 30782224

Wolfers T, Rokicki J, Alnaes D, Berthet P, Agartz I, Kia SM, Kaufmann T, Zabihi M, Moberget T, Melle I, Beckmann CF, Andreassen OA, Marquand AF, Westlye LT. 2021. Replicating extensive brain structural heterogeneity in individuals with schizophrenia and bipolar disorder. *Human Brain Mapping* **42**:2546–2555. DOI: https://doi.org/10.1002/hbm.25386, PMID: 33638594

Woo CW, Schmidt L, Krishnan A, Jepma M, Roy M, Lindquist MA, Atlas LY, Wager TD. 2017. Quantifying cerebral contributions to pain beyond nociception. *Nature Communications* **8**:14211. DOI: https://doi.org/10.1038/ncomms14211, PMID: 28195170

Yeo BTT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zöllei L, Polimeni JR, Fischl B, Liu H, Buckner RL. 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology* **106**:1125–1165. DOI: https://doi.org/10.1152/jn.00338.2011, PMID: 21653723

Yu Q, Allen EA, Sui J, Arbabshirani MR, Pearlson G, Calhoun VD. 2012. Brain connectivity networks in schizophrenia underlying resting state functional magnetic resonance imaging. *Current Topics in Medicinal Chemistry* **12**:2415–2425. DOI: https://doi.org/10.2174/156802612805289890, PMID: 23279180

Zabihi M, Oldehinkel M, Wolfers T, Frouin V, Goyard D, Loth E, Charman T, Tillmann J, Banaschewski T, Dumas G, Holt R, Baron-Cohen S, Durston S, Bölte S, Murphy D, Ecker C, Buitelaar JK, Beckmann CF, Marquand AF.

2019. Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging* **4**:567–578. DOI: https://doi.org/10.1016/j.bpsc.2018.11.013, PMID: 30799285

Zabihi M, Floris DL, Kia SM, Wolfers T, Tillmann J, Arenas AL, Moessnang C, Banaschewski T, Holt R, Baron-Cohen S, Loth E, Charman T, Bourgeron T, Murphy D, Ecker C, Buitelaar JK, Beckmann CF, Marquand A, EU-AIMS LEAP Group. 2020. Fractionating autism based on neuroanatomical normative modeling. *Translational Psychiatry* **10**:384. DOI: https://doi.org/10.1038/s41398-020-01057-0, PMID: 33159037

Zhang J, Kucyi A, Raya J, Nielsen AN, Nomi JS, Damoiseaux JS, Greene DJ, Horovitz SG, Uddin LQ, Whitfield-Gabrieli S. 2021. What have we really learned from functional connectivity in clinical populations? *NeuroImage* **242**:118466. DOI: https://doi.org/10.1016/j.neuroimage.2021.118466, PMID: 34389443

Appendix 1

Functional MRI Acquisition Parameters

In the HCP study, four runs of resting state fMRI data (14.5 min each) were acquired on a Siemens Skyra 3 Tesla scanner using multi-band gradient-echo EPI (TR = 720ms, TE = 33ms, flip angle = 52°, multiband acceleration factor = 8, 2 mm isotropic voxels, FOV = 208 × 180 mm, 72 slices, alternating RL/LR phase encode direction). T1 weighted scans were acquired with 3D MPRAGE sequence (TR = 2400ms, TE = 2.14ms, TI = 1000ms, flip angle = 8, 0.7 mm isotropic voxels, FOV = 224 mm, 256 sagittal slices) and T2 weighted scans were acquired with a SPACE sequence (TR = 3200ms, TE = 565ms, 0.7 mm isotropic voxels, FOV = 224 mm, 256 sagittal slices). In the COBRE study, the T1 weighted acquisition is a multi-echo MPRAGE (MEMPR) sequence (1 mm isotropic). Resting state functional MRI data was collected with single-shot full k-space echo-planar imaging (EPI) (TR = 2000ms, TE = 29ms, FOV = 64 × 64, 32 slices in axial plane interleaved multi slice series ascending, voxel size = 3 × 3 x4 mm^3). The University of Michigan SchizGaze study was collected in two phases with different parameters but using the same MRI machine (3.0T GE MR 750 Discovery scanner). In SchizGaze1 (N=47), functional images were acquired with a T2*-weighted, reverse spiral acquisition sequence (TR = 2000ms, 240 volumes (8 min), 3 mm isotropic voxels) and a T1-weighted image was acquired in the same prescription as the functional images to facilitate co-registration. In SchizGaze2 (N=68), functional images were acquired with a T2*-weighted multi-band EPI sequence (multi-band acceleration factor of 8, TR = 800ms, 453 volumes (6 min), 2.4 mm isotropic voxels) and T1w (MPRAGE) and T2w structural scans were acquired for co-registration with the functional data. In addition, field maps were acquired to correct for intensity and geometric distortions.

Functional MRI Preprocessing Methods

T1w images are corrected for intensity nonuniformity, reconstructed with recon-all (FreeSurfer), spatially normalized (ANTs), and segmented with FAST (FSL). For every BOLD run, data are co-registered to the corresponding T1w reference, and the BOLD signal is sampled onto the subject's surfaces with mri_vol2surf (FreeSurfer). A set of noise regressors are generated during the preceding steps that are used to remove a number of artifactual signals from the data during subsequent processing, and these noise regressors include: head-motion parameters (via MCFLIFT; FSL) framewise displacement and DVARS, and physiological noise regressors for use in component-based noise correction (CompCor). ICA-based denoising is implemented via ICA-AROMA and we compute 'non-aggressive' noise regressors. Resting state connectomes are generated from the fMRIPrep processed resting state data using Nilearn, denoising using the noise regressors generated above, with orthogonalization of regressors to avoid reintroducing artifactual signals.

Functional Brain Networks Normative Modeling

Data from 40 sites were combined to create the initial full sample. These sites are described in detail in ***Supplementary file 1***, including the sample size, age (mean and standard deviation), and sex distribution of each site. Many sites were pulled from publicly available datasets including ABCD, CAMCAN, CMI-HBN, HCP-Aging, HCP-Development, HCP-Early Psychosis, HCP-Young Adult, NKI-RS, OpenNeuro, PNC, and UKBiobank. For datasets that include repeated visits (i.e. ABCD, UKBiobank), only the first visit was included. Full details regarding sample characteristics, diagnostic procedures and acquisition protocols can be found in the publications associated with each of the studies. Training and testing datasets (80/20) were created using scikit-learn's train_test_split function, stratifying on the site variable. To show generalizability of the models to new data not included in training, we leveraged three datasets (ds000243, ds002843, ds003798) from OpenNeuro.org to create a multi-site transfer dataset.

Normative modeling was run using python 3.8 and the PCNtoolkit package (version 0.26). Bayesian Linear Regression (BLR) with likelihood warping was used to predict each Yeo-17 network pair from a vector of covariates (age, sex, mean_FD, site). For a detailed mathematical description see ***Fraza et al., 2021***. Briefly, for each region of interest, y is predicted as:

$$y = w^T \phi\left(x\right) + \epsilon \tag{1}$$

Where w^T is the estimated weight vector, $\phi\left(x\right)$ is a basis expansion of the of covariate vector **x,** consisting of a B-spline basis expansion (cubic spline with 5 evenly spaced knots) to model non-linear effects of age, and $\epsilon = \eta\left(0, \beta\right)$ a Gaussian noise distribution with mean zero and noise precision term β (the inverse variance). A likelihood warping approach was used to model non-Gaussian effects by applying a bijective nonlinear warping function to the non-Gaussian response variables to map them to a Gaussian latent space where inference can be performed in closed form. We used a 'sinarcsinsh' warping function, equivalent to the SHASH distribution that is commonly used in the generalized additive modeling literature (**Jones and Pewsey, 2009**). Site variation was modeled using fixed effects. A fast numerical optimization algorithm was used to optimize hyperparameters ('Powell'). Deviation scores (Z-scores) are calculated for the *n-th* subject, and *d-th* brain area, in the test set as:

$$Z_{nd} = \frac{y_{nd} - \hat{y}_{nd}}{\sqrt{\sigma_d^2 + \left(\sigma_*^2\right)_d}} \qquad (2)$$

Where y_{nd} is the true response, $\hat{y}_{nd}$ is the predicted mean, σ_d^2 is the estimated noise variance (reflecting uncertainty in the data), and $\left(\sigma_*^2\right)_d$ is the variance attributed to modeling uncertainty. Model fit for each brain region was evaluated by calculating the explained variance (which measures central tendency), the mean squared log-loss (MSLL, central tendency and variance) plus skew and kurtosis of the deviation scores (2) which measures how well the shape of the regression function matches the data (**Dinga et al., 2021**).

All pretrained models and code are shared online including documentation for transferring to new sites and an example transfer dataset. Given a new set of data (e.g. sites not present in the training set), this is done by first applying the warp parameters estimating on the training data to the new dataset, adjusting the mean and variance in the latent Gaussian space, then (if necessary) warping the adjusted data back to the original space, which is similar to the approach outlined in **Dinga et al., 2021**.