Evidence for Embracing Normative Modeling

Saige Rutherford[1,2,3*], **Pieter Barkema**[1], **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*†]

***For correspondence:**
saige.rutherford@donders.ru.nl;
andre.marquand@donders.ru.nl

†These authors contributed equally to this work

[1]Donders Institute for Brain, Cognition, and Behavior, Radboud University, Nijmegen, the Netherlands; [2]Department of Cognitive Neuroscience, Radboud University Medical Center, Nijmegen, the Netherlands; [3]Department of Psychiatry, University of Michigan, Ann Arbor, MI, United States; [4]Department of Psychology, University of Michigan, Ann Arbor, MI, United States; [5]Department of Philosophy, University of Michigan, Ann Arbor, MI, United States; [6]Centre for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, United Kingdom; [7]Department of Psychiatry, Radboud University Medical Center, Nijmegen, the 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 confirm the advantage (i.e., stronger effect sizes, more accurate classification and prediction) of using normative modeling features. We intend for these accessible resources to facilitate wider adoption of normative modeling across the neuroimaging community.

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.*** (***2016***); ***Rutherford et al.*** (***2022b***); ***Marquand et al.*** (***2019***). 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 often just the first step first , 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 within pediatric medicine, where conventional growth charts are used to map 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.*** (***2020***); ***Lv et al.*** (***2020***); ***Wolfers et al.*** (***2015***, 2017, 2021, 2018); ***Zabihi et al.*** (***2019***, 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 framework, 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 and Alloy*** (***2021***); ***Nour et al.*** (***2022***); ***Sanislow*** (***2020***); ***Zhang et al.*** (***2021***). Correspondingly, traditional analysis techniques for modeling case versus control effects has often led to null findings ***Winter et al.*** (***2022***) or significant but very small clinically meaningless differences. These effects are furthermore frequently unspecific 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.*** (***2021***, 2022); ***Kia and Marquand*** (***2018***); ***Kia et al.*** (***2020***, 2021, 2018); ***Kumar*** (***2021***); ***Boer et al.*** (***2022***)). 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). 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.

Methods and Materials

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. The full details of the data included in the functional normative model training can be found in the supplement methods (and Table 3). We leverage several data sets (Table 1) for the bench-

Table 1. Data Set Inclusion and Sample Overview.

		Cortical Thickness			Functional Networks		
Study	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, 6.6	499	28.9, 3.6	54.3, 45.6
COBRE	Classification and Group Difference	124	37.0, 12.7	24.2, 75.8	121	35.4, 12.4	23.1, 76.9
UMich_SZG	Classification and Group Difference	89	32.6, 9.6	50.6, 49.3	87	33.0, 10.1	50.6, 49.3

marking 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 University of Michigan SchizGaze (UMich) ***Tso et al.*** (***2021***). 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 et al.*** (***2020***). The HCP data was used in the prediction – regression benchmarking task. The COBRE and UMich data sets are used in the classification and group difference testing benchmarking tasks. Inclusion criteria across all the datasets was 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 (Euler < 10). 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 minutes of low motion data (framewise displacement < 0.5mm). 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 supplement).

Demographic, Cognition, 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.*** (***2020***). For COBRE and UMich studies, clinical diagnosis of schizophrenia was confirmed using the Structured Clinical Interview used for DSM-5 disorders (SCID) ***First and Williams*** (***2016***). All subjects were screened and excluded if they had: 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 ***Linden*** (***2012***); ***Fischl and Dale*** (***2000***); ***Fischl et al.*** (***2002***) 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 supplemental methods and table 4).

Resting-state data were preprocessed separately for each study using fMRIPrep ***Esteban et al.*** (***2018***); 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 (2x voxel size, 4-6mm), and regression of nuisance confounders (WM/CSF signals, non-aggressive AROMA components ***Pruim et al.*** (***2015b***,a), 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 the supplemental materials.

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 was estimated on approximately 58,000 subjects using a warped Bayesian Linear Regression algorithm ***Fraza et al.*** (***2021***). 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 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. Models were trained on a large multi-site data set (approx. N=22,000) and evaluated in several test sets using explained variance, mean standardized log loss, skew, and kurtosis. We transferred the functional normative models to the data sets 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, raw 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 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. 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 in the evaluation section below. An overview of the analysis workflow is shown in Figure 1.

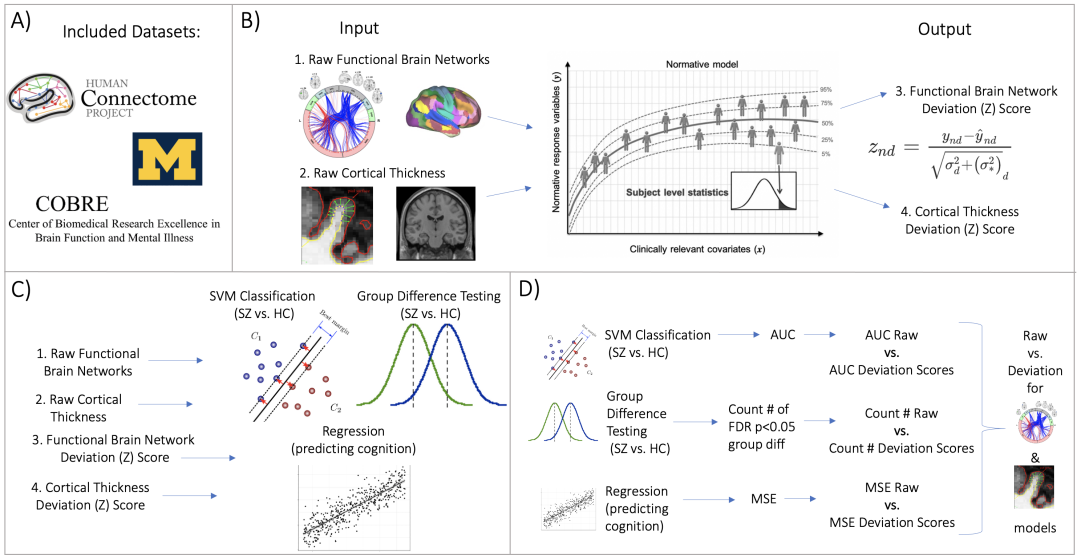

Figure 1. Figure 1 Overview of Workflow. **A)** Datasets included the Human Connectome Project (young adult) study, University of Michigan schizophrenia study, and COBRE schizophrenia study. **B)** Openly shared, pre-trained on big data, normative models were estimated for large scale resting state functional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

Task 1 Mass Univariate Group Difference Testing

Mass univariate group difference (schizophrenia vs. 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 correction, 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.

Task 2 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 the 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. 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***).

Task 3 Multivariate Prediction – Regression

A linear regression model was implemented to predict a latent variable of cognition (general cognitive ability) in the HCP data set. Brain Basis Set (BBS) is a predictive modeling approach developed and validated in previous studies ***Sripada et al.*** (***2019a***,b); see also studies by Wager and colleagues for a broadly similar approach ***Woo et al.*** (***2017***); ***Wager et al.*** (***2013***). BBS is similar to principal component regression, with an added predictive element. In the training set, PCA is performed on an $n_{subjects}$ x $p_{features}$ matrix using the PCA function from scikit-learn in Python, yielding components ordered by descending eigenvalues. Expression scores are then calculated for each of 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 $kx1$ 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***).

Evaluation

Evaluation for each benchmark task was done by estimating the appropriate model performance metric. For task one, the metric was the total count of models with significant group differences after multiple comparison correction (FDR-corrected $p < 0.05$). In task two, the metric was area under the receiving operator curve (AUC) averaged across all folds within a 10-fold cross validation framework. For task three, the metric was the mean squared error (MSE) of the prediction in the test set. Evaluation metrics of each task were calculated independently for both deviation score (Z) and raw data (R) models. Higher AUC, higher count, and lower MSE represent better model performance. We then have a statistic of interest that is observed, θ, which represents the difference between deviation and raw data model performance.

$$\theta_{task1} = Count_z - Count_R \quad (1)$$

$$\theta_{task2} = AUC_z - AUC_R \quad (2)$$

$$\theta_{task3} = MSE_R - MSE_z \quad (3)$$

To assess whether θ is more likely than would be expected by chance, we generated the null distribution for θ using permutations. Within one iteration of the permutation framework, a random sample is generated by shuffling the labels (In task 1 and 2 we shuffle the clinical group labels, and in task 3 we shuffle the g-factor 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 perm iteration). The shuffled models are evaluated, and we calculate θ_{perm} for each random shuffle of labels. We set $n_{perm} = 10,000$ and use the distribution of θ_{perm} to calculate a p-value for θ_{obs} at each benchmarking task. The permuted p-value is equal to $(C + 1)/(n_{perm} + 1)$. Where C is the number of permutations where $\theta_{perm} > \theta_{obsd}$. The same evaluation procedure described here (including permutations) was performed for both cortical thickness and functional network modalities.

Results

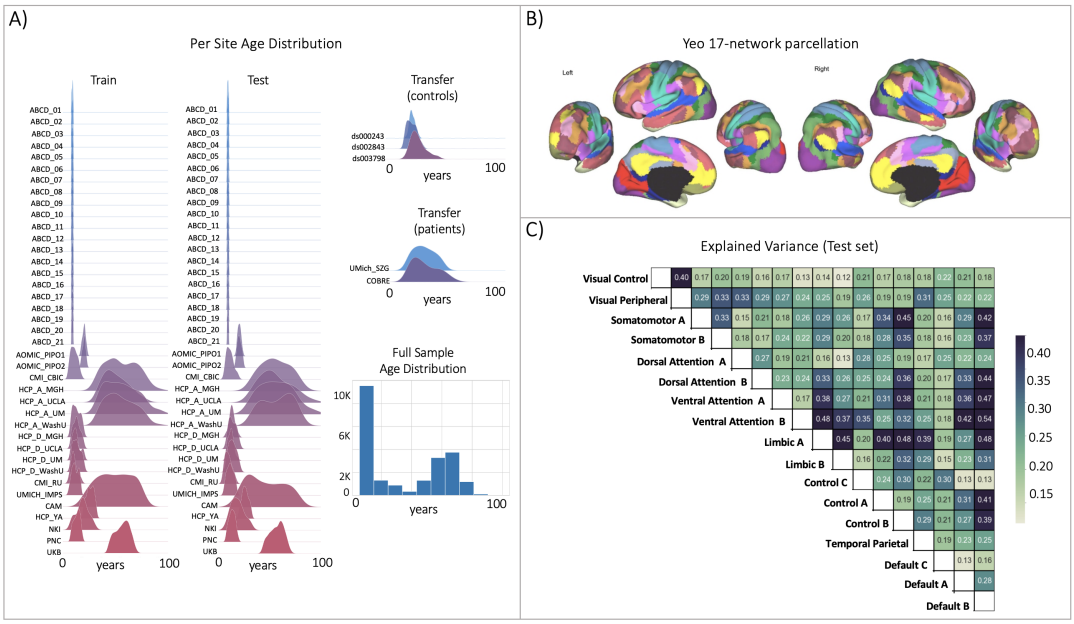

Figure 2. Functional brain network normative modeling. **A)** Age distribution per scanning site in all the train/test/transfer data partitions and across the full sample (train + test). **B)** The Yeo-17 brain network atlas used to generate connectomes Between network connectivity was calculated for 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.

Sharing of functional big data normative models

The first result of this work is the evaluation of the functional big data normative models (Figure 3). 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 data sets 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 as 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).

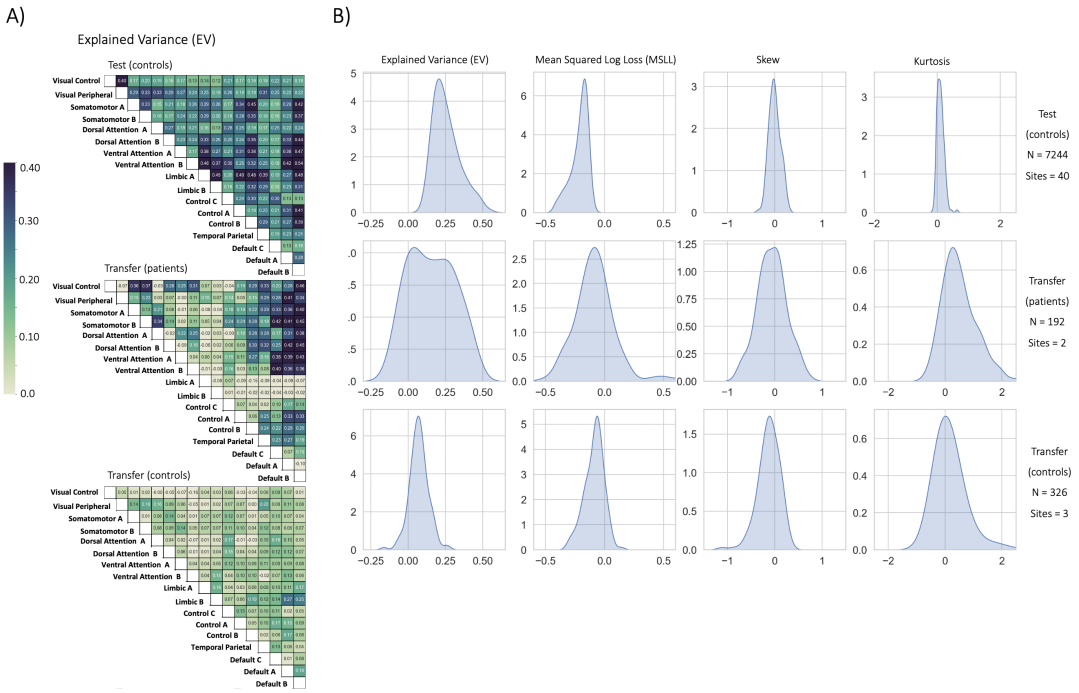

Figure 3. Functional Normative Model Evaluation Metrics. **A)** Explained variance per network pair across the test set (top), and both transfer sets (patients – middle, controls – bottom).We point out that the age range of the transfer (controls) sample (shown in Figure 2A) falls into a range with sparse data, and therefore the lower explained variance observed in the transfer (controls) group compared to the test and transfer (patients) groups is likely due to epistemic uncertainty (reducible with adding more data points) of the model predictions in this age range. **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 1), more negative MSLL, and normally distributed skew and kurtosis correspond to better model fit.

Benchmarking Task One Mass Univariate Group Difference Testing

The strongest evidence for embracing normative modeling can be seen in the benchmarking task one group difference (schizophrenia vs. controls) testing results (Table 2, Figure 4A). In this application, we observe numerous group differences in both functional and structural deviation score models after applying stringent multiple comparison correction (FDR $p-value < 0.05$). The strongest effects ($HC > SZ$) in the structural models were located in the right hemisphere lateral occipito-temporal 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 Sensorimotor B-Default B, Dorsal Attention B-Default B, Sensorimotor B-Default A, Control B-Default A, and Ventral Attention A-Default B. In the raw data models, which were residualized of covariates including site, sex, and linear plus quadratic effects of age and head motion (only included for functional models), we observe no group differences after multiple comparison correction. 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***, 2016); ***Dansereau et al.*** (***2017***); ***Howes et al.*** (***2022***); ***Lei***

Table 2. Benchmarking Results. Deviation (Z) score column shows the performance using deviation scores (AUC for classification, total number of regions with significant group differences FDR-corrected p<0.05 for case vs. 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 better performance using the deviation scores. * = statistically significant difference between Z and Raw established using permutation testing (10k perms).

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

et al. (*2020b*,a); *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 investigation of the uncorrected statistical maps, we observe that the raw data follows a similar pattern to the deviation group difference map, but these results do not withstand multiple comparison correction.

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 reveal subtle individual differences which are lost when only looking at group means. The individual difference maps shows 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 ($p = 0.0067$).

Benchmarking Task 2 Multivariate Prediction – Classification

In benchmarking task two, we classified schizophrenia versus controls using support vector classification 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.

Benchmarking Task 3 Multivariate Prediction – Regression

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

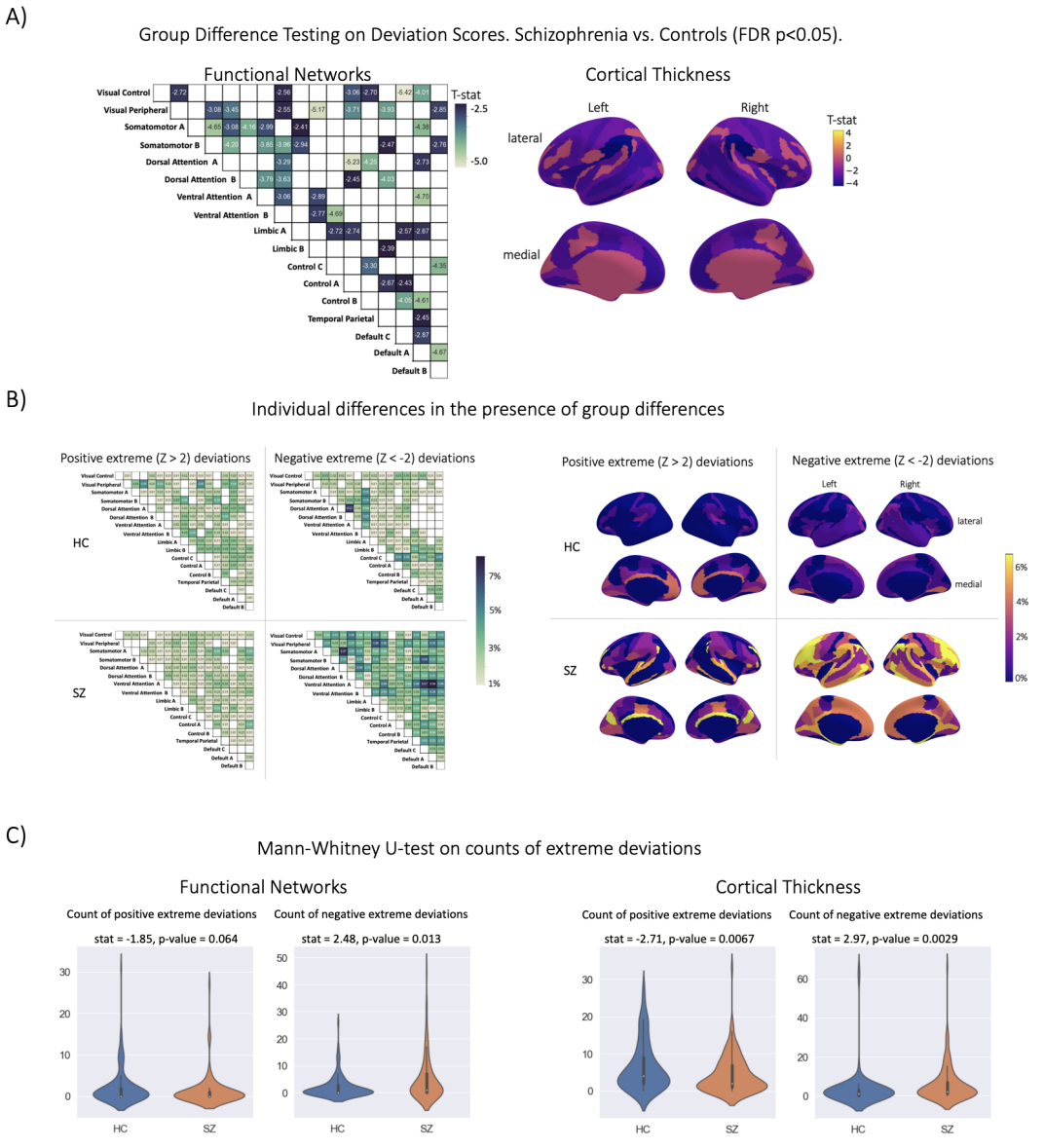

Figure 4. Group Difference Testing Evaluation. **A)** Significant group differences in the deviation score models, (top left) functional brain network deviation and (top right) cortical thickness deviation scores. The raw data, either cortical thickness or functional brain networks (residualized of sex and linear/ quadratic effects of age and motion (mean framewise displacement)) resulted in no significant group differences after multiple comparison correction. **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. **D)** There are significant group differences in the summaries (count) of the individual difference maps (panel B).

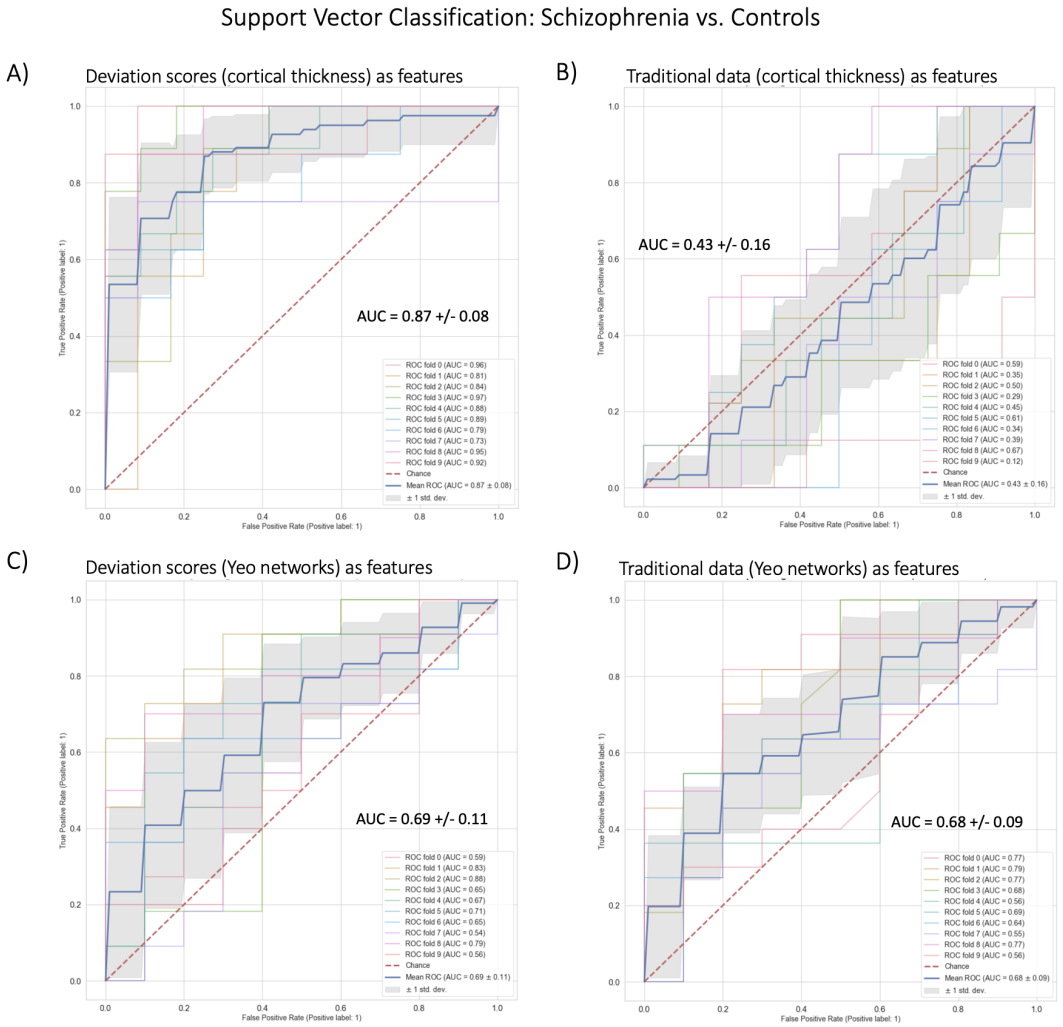

Figure 5. Benchmark Task 2 Multivariate Prediction – Classification Evaluation. **A)** Support Vector Classification using cortical thickness deviation scores as input features (most accurate model). **B)** Support Vector Classification using cortical thickness (residualized of sex and linear/quadratic effects of age) as input features. **C)** Support Vector Classification using functional brain network deviation scores as input features. **D)** Support Vector Classification using functional brain networks (residualized of sex and linear/ quadratic effects of age and motion (mean framewise displacement)) as input features.

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 data set combined with a user-friendly interface. 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 versus control), multivariate prediction – classification (schizophrenia versus 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 confirm 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.

Prior works on the methodological innovation and application of normative modeling ***Kia et al.*** (***2018***); ***Kia and Marquand*** (***2018***); ***Kia et al.*** (***2020***, 2021); ***Boer et al.*** (***2022***)) 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 versus control approaches). This is due in part to the challenges arising from the results existing in a very high dimensional space (100s to 1000s of brain regions from 100s to 1000s 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 superiority of normative modeling over case-control methods. Rather, our results, especially in the comparisons of group difference maps to individual difference maps (Figure 4), from this work 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 to 1,000 nodes and 5,000-500,000 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***); ***Kong et al.*** (***2019***); ***Laumann et al.*** (***2015***); ***Power et al.*** (***2011***); ***Schaefer et al.*** (***2018***); ***Shen et al.*** (***2013***). There is 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 ***Smith et al.*** (***2009***), though we 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 refers curious readers to a thorough review on 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 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 which 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 ***Colyvan*** (***2013***); ***Baron*** (***2004***); ***Catita et al.*** (***2020***). 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 towards 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 of 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 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 work 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.

Code and Data Availability

Pre-trained normative models are available on GitHub and Google Colab. Scripts for running the benckmarking analysis and visualizations are available on GitHub here. Online portal for running

models without code is in beta testing phase and will be available here shortly.

Acknowledgments and Contributions

Conception of the work - SR. Data curation and management - SR. Data analysis and interpretation - SR, PB. Writing the article - SR. Revision of the article - PB, IFT, CS, CFB, HG, AFM. Supervision and Funding - IFT, CS, CFB, HG, AFM. This research was supported by grants from the European Research Council (ERC, grant "MENTALPRECISION" 10100118 and "BRAINMINT" 802998), the Wellcome Trust under an Innovator award ("BRAINCHART", 215698/Z/19/Z) and a Strategic Award (098369/Z/12/Z), the Dutch Organisation for Scientific Research (VIDI grant 016.156.415). IFT was funded by National Institute of Mental Health K23MH108823. CS was funded by the National Institute of Mental Health R01MH107741.

Conflicts of Interest

CFB is director and shareholder of SBGNeuro Ltd. HGR received speaker's honorarium from Lundbeck and Janssen. The other authors report no conflicts of interest.

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. Multimodal Neuroimaging in Schizophrenia: Description and Dissemination. Neuroinformatics. 2017; 15(4):343–364. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5671541/, doi: 10.1007/s12021-017-9338-9.

Arbabshirani MR, Kiehl KA, Pearlson GD, Calhoun VD. Classification of schizophrenia patients based on resting-state functional network connectivity. Frontiers in Neuroscience. 2013; 7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3744823/, doi: 10.3389/fnins.2013.00133.

Baker JT, Dillon DG, Patrick LM, Roffman JL, Brady RO, Pizzagalli DA, Öngür D, Holmes AJ. Functional connectomics of affective and psychotic pathology. Proceedings of the National Academy of Sciences. 2019; 116(18):9050–9059. https://www.pnas.org/doi/abs/10.1073/pnas.1820780116, doi: 10.1073/pnas.1820780116, publisher: Proceedings of the National Academy of Sciences.

Baron J. Normative Models of Judgment and Decision Making. In: Koehler DJ, Harvey N, editors. *Blackwell Handbook of Judgment and Decision Making* Blackwell Publishing Ltd; 2004.p. 19–36. https://onlinelibrary.wiley.com/doi/10.1002/9780470752937.ch2, doi: 10.1002/9780470752937.ch2.

Benkarim O, Paquola C, Park By, Kebets V, Hong SJ, Wael RVd, Zhang S, Yeo BTT, Eickenberg M, Ge T, Poline JB, Bernhardt BC, Bzdok D. Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. PLOS Biology. 2022; 20(4):e3001627. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001627, doi: 10.1371/journal.pbio.3001627, publisher: Public Library of Science.

Bijsterbosch J, Harrison SJ, Jbabdi S, Woolrich M, Beckmann C, Smith S, Duff EP. Challenges and future directions for representations of functional brain organization. Nature Neuroscience. 2020; 23(12):1484–1495. https://www.nature.com/articles/s41593-020-00726-z, doi: 10.1038/s41593-020-00726-z, number: 12 Publisher: Nature Publishing Group.

Boer AAAd, Kia SM, Rutherford S, Zabihi M, Fraza C, Barkema P, Westlye LT, Andreassen OA, Hinne M, Beckmann CF, Marquand A. Non-Gaussian Normative Modelling With Hierarchical Bayesian Regression. bioRxiv. 2022; https://www.biorxiv.org/content/10.1101/2022.10.05.510988v1, doi: 10.1101/2022.10.05.510988.

Borghi E, Onis Md, Garza C, Broeck JVd, Frongillo EA, Grummer-Strawn L, Buuren SV, Pan H, Molinari L, Martorell R, Onyango AW, Martines JC. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. Statistics in Medicine. 2006; 25(2):247–265. http://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2227, doi: 10.1002/sim.2227, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2227.

Cai N, Choi KW, Fried EI. Reviewing the genetics of heterogeneity in depression: operationalizations, manifestations and etiologies. Human Molecular Genetics. 2020; 29:R10–R18. https://academic.oup.com/hmg/article/29/R1/R10/5860824, doi: 10.1093/hmg/ddaa115.

480 **Catita M**, Águas A, Morgado P. Normality in medicine: a critical review. Philosophy, Ethics, and Humanities
481 in Medicine. 2020; 15(1):3. https://peh-med.biomedcentral.com/articles/10.1186/s13010-020-00087-2, doi:
482 10.1186/s13010-020-00087-2.

483 **Cetin MS**, Houck JM, Rashid B, Agacoglu O, Stephen JM, Sui J, Canive J, Mayer A, Aine C, Bustillo JR, Calhoun
484 VD. Multimodal Classification of Schizophrenia Patients with MEG and fMRI Data Using Static and Dynamic
485 Connectivity Measures. Frontiers in Neuroscience. 2016; 10. https://www.ncbi.nlm.nih.gov/pmc/articles/
486 PMC5070283/, doi: 10.3389/fnins.2016.00466.

487 **Cetin MS**, Houck JM, Vergara VM, Miller RL, Calhoun V. Multimodal Based Classification of Schizophrenia Pa-
488 tients. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and
489 Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2015; 2015:2629–2632.
490 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4880008/, doi: 10.1109/EMBC.2015.7318931.

491 **Colyvan M**. Idealisations in normative models. Synthese. 2013; 190(8):1337–1350. https://doi.org/10.1007/
492 s11229-012-0166-z, doi: 10.1007/s11229-012-0166-z.

493 **Cuthbert BN**, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Medicine.
494 2013; 11:126. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3653747/, doi: 10.1186/1741-7015-11-126.

495 **Dansereau C**, Benhajali Y, Risterucci C, Pich EM, Orban P, Arnold D, Bellec P. Statistical power and prediction ac-
496 curacy in multisite resting-state fMRI connectivity. NeuroImage. 2017; 149:220–232. http://www.sciencedirect.
497 com/science/article/pii/S1053811917300939, doi: 10.1016/j.neuroimage.2017.01.072.

498 **Dinga R**, Fraza CJ, Bayer JMM, Kia SM, Beckmann CF, Marquand AF. Normative modeling of neuroimaging data
499 using generalized additive models of location scale and shape. bioRxiv. 2021; http://biorxiv.org/lookup/doi/
500 10.1101/2021.06.14.448106, doi: 10.1101/2021.06.14.448106.

501 **Dong HM**, Castellanos FX, Yang N, Zhang Z, Zhou Q, He Y, Zhang L, Xu T, Holmes AJ, Thomas Yeo BT, Chen F, Wang
502 B, Beckmann C, White T, Sporns O, Qiu J, Feng T, Chen A, Liu X, Chen X, et al. Charting brain growth in tandem
503 with brain templates at school age. Science Bulletin. 2020; 65(22):1924–1934. https://www.sciencedirect.com/
504 science/article/pii/S2095927320304965, doi: 10.1016/j.scib.2020.07.027.

505 **Eickhoff SB**, Yeo BTT, Genon S. Imaging-based parcellations of the human brain. Nature Reviews Neuroscience.
506 2018; 19(11):672–686. https://www.nature.com/articles/s41583-018-0071-7, doi: 10.1038/s41583-018-0071-7,
507 number: 11 Publisher: Nature Publishing Group.

508 **van Erp TGM**, Walton E, Hibar DP, Schmaal L, Jiang W, Glahn DC, Pearlson GD, Yao N, Fukunaga M, Hashimoto
509 R, Okada N, Yamamori H, Bustillo JR, Clark VP, Agartz I, Mueller BA, Cahn W, de Zwarte SMC, Hulshoff Pol
510 HE, Kahn RS, et al. Cortical Brain Abnormalities in 4474 Individuals With Schizophrenia and 5098 Control
511 Subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium. Biological
512 Psychiatry. 2018; 84(9):644–654. doi: 10.1016/j.biopsych.2018.04.023.

513 **Esteban O**, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M,
514 Oya H, Ghosh SS, Wright J, Durnez J, Poldrack RA, Gorgolewski KJ. fMRIPrep: a robust preprocessing pipeline
515 for functional MRI. Nature Methods. 2018; p. 1. https://www.nature.com/articles/s41592-018-0235-4, doi:
516 10.1038/s41592-018-0235-4.

517 **Etkin A**. A Reckoning and Research Agenda for Neuroimaging in Psychiatry. American Journal of Psy-
518 chiatry. 2019; 176(7):507–511. https://ajp.psychiatryonline.org/doi/10.1176/appi.ajp.2019.19050521, doi:
519 10.1176/appi.ajp.2019.19050521, publisher: American Psychiatric Publishing.

520 **Filip P**, Bednarik P, Eberly LE, Moheet A, Svatkova A, Grohn H, Kumar AF, Seaquist ER, Mangia S. Different
521 FreeSurfer versions might generate different statistical outcomes in case–control comparison studies. Neu-
522 roradiology. 2022; 64(4):765–773. https://doi.org/10.1007/s00234-021-02862-0, doi: 10.1007/s00234-021-
523 02862-0.

524 **Finn ES**, Rosenberg MD. Beyond fingerprinting: Choosing predictive connectomes over reliable connectomes.
525 NeuroImage. 2021; 239:118254. https://www.sciencedirect.com/science/article/pii/S1053811921005310, doi:
526 10.1016/j.neuroimage.2021.118254.

527 **First MB**, Williams BW. SCID-5-CV: structured clinical interview for DSM-5 disorders: clinician version. American
528 Psychiatric Association Publishing,; 2016. Publication Title: SCID-5-CV: structured clinical interview for DSM-5
529 disorders: clinician version.

530 **Fischl B**, Dale AM. Measuring the thickness of the human cerebral cortex from magnetic resonance im-
531 ages. Proc Natl Acad Sci U S A. 2000; 97(20):11050–11055. http://dx.doi.org/10.1073/pnas.200033797, doi:
532 10.1073/pnas.200033797.

533 **Fischl B**, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness
534 S, Montillo A, Makris N, Rosen B, Dale AM. Whole Brain Segmentation. Neuron. 2002; 33(3):341–355. http:
535 //dx.doi.org/10.1016/s0896-6273(02)00569-x, doi: 10.1016/s0896-6273(02)00569-x.

536 **Flake JK**, Fried EI. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid
537 Them. Advances in Methods and Practices in Psychological Science. 2020; p. 2515245920952393. https:
538 //doi.org/10.1177/2515245920952393, doi: 10.1177/2515245920952393, publisher: SAGE Publications Inc.

539 **Floris DL**, Wolfers T, Zabihi M, Holz NE, Zwiers MP, Charman T, Tillmann J, Ecker C, Dell'Acqua F, Banaschewski T,
540 Moessnang C, Baron-Cohen S, Holt R, Durston S, Loth E, Murphy DGM, Marquand A, Buitelaar JK, Beckmann
541 CF, Ahmad J, et al. Atypical Brain Asymmetry in Autism—A Candiyear for Clinically Meaningful Stratifica-
542 tion. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. 2020; https://www.sciencedirect.com/
543 science/article/pii/S2451902220302433, doi: 10.1016/j.bpsc.2020.08.008.

544 **Fraza C**, Zabihi M, Beckmann CF, Marquand AF, The Extremes of Normative Modelling. bioRxiv; 2022.
545 https://www.biorxiv.org/content/10.1101/2022.08.23.505049v1, doi: 10.1101/2022.08.23.505049, pages:
546 2022.08.23.505049 Section: New Results.

547 **Fraza CJ**, Dinga R, Beckmann CF, Marquand AF. Warped Bayesian linear regression for normative mod-
548 elling of big data. NeuroImage. 2021; 245:118715. https://www.sciencedirect.com/science/article/pii/
549 S1053811921009873, doi: 10.1016/j.neuroimage.2021.118715.

550 **Gau R**, Noble S, Heuer K, Bottenhorn KL, Bilgin IP, Yang YF, Huntenburg JM, Bayer JMM, Bethlehem RAI,
551 Rhoads SA, Vogelbacher C, Borghesani V, Levitis E, Wang HT, Van Den Bossche S, Kobeleva X, Legarreta
552 JH, Guay S, Atay SM, Varoquaux GP, et al. Brainhack: Developing a culture of open, inclusive, community-
553 driven neuroscience. Neuron. 2021; 109(11):1769–1775. https://www.sciencedirect.com/science/article/pii/
554 S0896627321002312, doi: 10.1016/j.neuron.2021.04.001.

555 **Gershon RC**, Cella D, Fox NA, Havlik RJ, Hendrie HC, Wagster MV. Assessment of neurological and behavioural
556 function: the NIH Toolbox. The Lancet Neurology. 2010; 9(2):138–139. http://www.sciencedirect.com/science/
557 article/pii/S1474442209703357, doi: 10.1016/S1474-4422(09)70335-7.

558 **Glasser MF**, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF,
559 Jenkinson M. A multi-modal parcellation of human cerebral cortex. Nature. 2016; 536(7615):171–178.

560 **Goodkind M**, Eickhoff SB, Oathes DJ, Jiang Y, Chang A, Jones-Hagata LB, Ortega BN, Zaiko YV, Roach EL, Kor-
561 gaonkar MS, Grieve SM, Galatzer-Levy I, Fox PT, Etkin A. Identification of a Common Neurobiological Sub-
562 strate for Mental Illness. JAMA Psychiatry. 2015; 72(4):305–315. https://doi.org/10.1001/jamapsychiatry.2014.
563 2206, doi: 10.1001/jamapsychiatry.2014.2206.

564 **Greene AS**, Shen X, Noble S, Horien C, Hahn CA, Arora J, Tokoglu F, Spann MN, Carrión CI, Barron DS, Sanacora
565 G, Srihari VH, Woods SW, Scheinost D, Constable RT. Brain–phenotype models fail for individuals who defy
566 sample stereotypes. Nature. 2022; 609(7925):109–118. https://www.nature.com/articles/s41586-022-05118-w,
567 doi: 10.1038/s41586-022-05118-w, number: 7925 Publisher: Nature Publishing Group.

568 **Henrich J**, Heine SJ, Norenzayan A. The weirdest people in the world? Behavioral and Brain Sci-
569 ences. 2010; 33(2):61–83. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/
570 abs/weirdest-people-in-the-world/BF84F7517D56AFF7B7EB58411A554C17, doi: 10.1017/S0140525X0999152X,
571 publisher: Cambridge University Press.

572 **Holz NE**, Floris DL, Llera A, Aggensteiner PM, Kia SM, Wolfers T, Baumeister S, Böttinger B, Glennon JC, Hoekstra
573 PJ, Dietrich A, Saam MC, Schulze UME, Lythgoe DJ, Williams SCR, Santosh P, Rosa-Justicia M, Bargallo N, Castro-
574 Fornieles J, Arango C, et al. Age-related brain deviations and aggression. Psychological Medicine. 2022; p.
575 1–10. doi: 10.1017/S003329172200068X.

576 **Howes OD**, Cummings C, Chapman GE, Shatalina E. Neuroimaging in schizophrenia: an overview of findings
577 and their implications for synaptic changes. Neuropsychopharmacology. 2022; p. 1–17. https://www.nature.
578 com/articles/s41386-022-01426-x, doi: 10.1038/s41386-022-01426-x, publisher: Nature Publishing Group.

579 **Insel T**, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, Sanislow C, Wang P. Research Domain Criteria
580 (RDoC): Toward a New Classification Framework for Research on Mental Disorders. American Journal of
581 Psychiatry. 2010; 167(7):748–751. http://ajp.psychiatryonline.org/doi/10.1176/appi.ajp.2010.09091379, doi:
582 10.1176/appi.ajp.2010.09091379, publisher: American Psychiatric Publishing.

Itälinna V, Kaltiainen H, Forss N, Liljeström M, Parkkonen L, Detecting mild traumatic brain injury with MEG, normative modelling and machine learning. medRxiv; 2022. https://www.medrxiv.org/content/10.1101/2022.09.29.22280521v1, doi: 10.1101/2022.09.29.22280521, pages: 2022.09.29.22280521.

Jones MC, Pewsey A. Sinh-arcsinh distributions. Biometrika. 2009; 96(4):761–780. https://www.jstor.org/stable/27798865.

Kia SM, Beckmann CF, Marquand AF. Scalable Multi-Task Gaussian Process Tensor Regression for Normative Modeling of Structured Variation in Neuroimaging Data. arXiv:180800036 [cs, stat]. 2018; http://arxiv.org/abs/1808.00036.

Kia SM, Huijsdens H, Dinga R, Wolfers T, Mennes M, Andreassen OA, Westlye LT, Beckmann CF, Marquand AF. Hierarchical Bayesian Regression for Multi-site Normative Modeling of Neuroimaging Data. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* Lecture Notes in Computer Science, Springer International Publishing; 2020. p. 699–709. doi: 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. Federated Multi-Site Normative Modeling using Hierarchical Bayesian Regression. bioRxiv. 2021; p. 2021.05.28.446120. https://www.biorxiv.org/content/10.1101/2021.05.28.446120v1, doi: 10.1101/2021.05.28.446120, publisher: Cold Spring Harbor Laboratory Section: New Results.

Kia SM, Marquand A. Normative Modeling of Neuroimaging Data using Scalable Multi-Task Gaussian Processes. arXiv:180601047 [cs, stat]. 2018; http://arxiv.org/abs/1806.01047.

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. Mapping Normative Trajectories of Cognitive Function and Its Relation to Psychopathology Symptoms and Genetic Risk in Youth. Biological Psychiatry Global Open Science. 2022; https://www.sciencedirect.com/science/article/pii/S266717432200012X, doi: 10.1016/j.bpsgos.2022.01.007.

Klapwijk ET, van de Kamp F, van der Meulen M, Peters S, Wierenga LM. Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data. NeuroImage. 2019; 189:116–129. doi: 10.1016/j.neuroimage.2019.01.014.

Kong R, Li J, Orban C, Sabuncu MR, Liu H, Schaefer A, Sun N, Zuo XN, Holmes AJ, Eickhoff SB, Yeo BTT. Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. Cerebral Cortex (New York, NY). 2019; 29(6):2533–2551. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519695/, doi: 10.1093/cercor/bhy123.

Kumar S. NormVAE: Normative Modeling on Neuroimaging Data using Variational Autoencoders. aRxiv. 2021; https://arxiv.org/abs/2110.04903v2.

Laumann T, Gordon E, Adeyemo B, Snyder A, Joo S, Chen MY, Gilmore A, McDermott K, Nelson S, Dosenbach NF, Schlaggar B, Mumford J, Poldrack R, Petersen S. Functional System and Areal Organization of a Highly Sampled Individual Human Brain. Neuron. 2015; 87(3):657–670. https://www.sciencedirect.com/science/article/pii/S0896627315006005, doi: 10.1016/j.neuron.2015.06.037.

Lee W, Bindman J, Ford T, Glozier N, Moran P, Stewart R, Hotopf M. Bias in psychiatric case–control studies: Literature survey. The British Journal of Psychiatry. 2007; 190(3):204–209. https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/bias-in-psychiatric-casecontrol-studies/CA1B1AD515086BE2E546BEDDD8E77614, doi: 10.1192/bjp.bp.106.027250, publisher: Cambridge University Press.

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. 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. 2020; 50(11):1852–1861. doi: 10.1017/S0033291719001934.

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. Integrating machining learning and multimodal neuroimaging to detect schizophrenia at the level of the individual. Human Brain Mapping. 2020; 41(5):1119–1135. doi: 10.1002/hbm.24863.

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. Centering inclusivity in the design of online conferences—An OHBM–Open Science perspective. GigaScience. 2021; 10(8):giab051. https://doi.org/10.1093/gigascience/giab051, doi: 10.1093/gigascience/giab051.

Li J, Bzdok D, Chen J, Tam A, Ooi LQR, Holmes AJ, Ge T, Patil KR, Jabbi M, Eickhoff SB, Yeo BTT, Genon S. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. Science Advances. 2022; 8(11):eabj1812. https://www.science.org/doi/full/10.1126/sciadv.abj1812, doi: 10.1126/sciadv.abj1812, publisher: American Association for the Advancement of Science.

Linden DEJ. The Challenges and Promise of Neuroimaging in Psychiatry. Neuron. 2012; 73(1):8–22. https://www.cell.com/neuron/abstract/S0896-6273(11)01095-6, doi: 10.1016/j.neuron.2011.12.014, publisher: Elsevier.

Loth E, Ahmad J, Chatham C, López B, Carter B, Crawley D, Oakley B, Hayward H, Cooke J, Cáceres ASJ, Bzdok D, Jones E, Charman T, Beckmann C, Bourgeron T, Toro R, Buitelaar J, Murphy D, Dumas G. The meaning of significant mean group differences for biomarker discovery. PLOS Computational Biology. 2021; 17(11):e1009477. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009477, doi: 10.1371/journal.pcbi.1009477, publisher: Public Library of Science.

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. Individual deviations from normative models of brain structure in a large cross-sectional schizophrenia cohort. Molecular Psychiatry. 2020; p. 1–12. https://www.nature.com/articles/s41380-020-00882-5, doi: 10.1038/s41380-020-00882-5, publisher: Nature Publishing Group.

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. Reproducible brain-wide association studies require thousands of individuals. Nature. 2022; p. 1–7. https://www.nature.com/articles/s41586-022-04492-9, doi: 10.1038/s41586-022-04492-9, publisher: Nature Publishing Group.

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

Marquand AF, Haak KV, Beckmann CF. Functional corticostriatal connection topographies predict goal directed behaviour in humans. Nature human behaviour. 2017; 1(8). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5549843/, doi: 10.1038/s41562-017-0146.

Marquand AF, Kia SM, Zabihi M, Wolfers T, Buitelaar JK, Beckmann CF. Conceptualizing mental disorders as deviations from normative functioning. Molecular Psychiatry. 2019; 24(10):1415–1424. https://www.nature.com/articles/s41380-019-0441-1, doi: 10.1038/s41380-019-0441-1, number: 10 Publisher: Nature Publishing Group.

Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. Biological Psychiatry. 2016; 80(7):552–561. https://www.sciencedirect.com/science/article/pii/S0006322316000020, doi: 10.1016/j.biopsych.2015.12.023.

McTeague LM, Huemer J, Carreon DM, Jiang Y, Eickhoff SB, Etkin A. Identification of Common Neural Circuit Disruptions in Cognitive Control Across Psychiatric Disorders. American Journal of Psychiatry. 2017; 174(7):676–685. https://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.2017.16040400, doi: 10.1176/appi.ajp.2017.16040400, publisher: American Psychiatric Publishing.

Meng X, Jiang R, Lin D, Bustillo J, Jones T, Chen J, Yu Q, Du Y, Zhang Y, Jiang T, Sui J, Calhoun VD. Predicting individualized clinical measures by a generalized prediction framework and multimodal fusion of MRI data. NeuroImage. 2017; 145:218–229. http://www.sciencedirect.com/science/article/pii/S105381191630146X, doi: 10.1016/j.neuroimage.2016.05.026.

Michelini G, Palumbo IM, DeYoung CG, Latzman RD, Kotov R. Linking RDoC and HiTOP: A new interface for advancing psychiatric nosology and neuroscience. Clinical Psychology Review. 2021; 86:102025. https://www.sciencedirect.com/science/article/pii/S0272735821000684, doi: 10.1016/j.cpr.2021.102025.

Moriarity DP, Alloy LB. Back to Basics: The Importance of Measurement Properties in Biological Psychiatry. Neuroscience & Biobehavioral Reviews. 2021; 123:72–82. https://www.sciencedirect.com/science/article/pii/S0149763421000221, doi: 10.1016/j.neubiorev.2021.01.008.

Mottron L, Bzdok D. Diagnosing as autistic people increasingly distant from prototypes lead neither to clinical benefit nor to the advancement of knowledge. Molecular Psychiatry. 2022; 27(2):773–775. https://www.nature.com/articles/s41380-021-01343-3, doi: 10.1038/s41380-021-01343-3, number: 2 Publisher: Nature Publishing Group.

Nour MM, Liu Y, Dolan RJ. Functional neuroimaging in psychiatry and the case for failing better. Neuron. 2022; 110(16):2524–2544. https://www.cell.com/neuron/abstract/S0896-6273(22)00647-X, doi: 10.1016/j.neuron.2022.07.005, publisher: Elsevier.

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 E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12(85):2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

Pereira-Sanchez V, Castellanos FX. Neuroimaging in attention-deficit/hyperactivity disorder. Current Opinion in Psychiatry. 2021; 34(2):105–111. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7879851/, doi: 10.1097/YCO.0000000000000669.

Power J, Cohen A, Nelson S, Wig G, Barnes K, Church J, Vogel A, Laumann T, Miezin F, Schlaggar B, Petersen S. Functional Network Organization of the Human Brain. Neuron. 2011; 72(4):665–678. http://www.sciencedirect.com/science/article/pii/S0896627311007926, doi: 10.1016/j.neuron.2011.09.006.

Pruim RHR, Mennes M, van Rooij D, Llera A, Buitelaar JK, Beckmann CF. ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. NeuroImage. 2015; 112:267–277. http://www.sciencedirect.com/science/article/pii/S1053811915001822, doi: 10.1016/j.neuroimage.2015.02.064.

Pruim RHR, Mennes M, Buitelaar JK, Beckmann CF. Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. NeuroImage. 2015; 112:278–287. http://linkinghub.elsevier.com/retrieve/pii/S1053811915001809, doi: 10.1016/j.neuroimage.2015.02.063.

Rahim M, Thirion B, Bzdok D, Buvat I, Varoquaux G. Joint prediction of multiple scores captures better individual traits from brain images. NeuroImage. 2017; 158:145–154. http://www.sciencedirect.com/science/article/pii/S1053811917305438, doi: 10.1016/j.neuroimage.2017.06.072.

Rios G, Tobar F. Compositionally-warped Gaussian processes. Neural Networks. 2019; 118:235–246. https://linkinghub.elsevier.com/retrieve/pii/S0893608019301856, doi: 10.1016/j.neunet.2019.06.012.

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

Rosenberg MD, Finn ES. How to establish robust brain–behavior relationships without thousands of individuals. Nature Neuroscience. 2022; 25(7):835–837. https://www.nature.com/articles/s41593-022-01110-9, doi: 10.1038/s41593-022-01110-9, number: 7 Publisher: Nature Publishing Group.

Rutherford S, Angstadt M, Sripada C, Chang SE. Leveraging big data for classification of children who stutter from fluent peers. bioRxiv. 2020; p. 2020.10.28.359711. https://www.biorxiv.org/content/10.1101/2020.10.28.359711v1, doi: 10.1101/2020.10.28.359711, publisher: Cold Spring Harbor Laboratory Section: New Results.

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. Charting brain growth and aging at high spatial precision. eLife. 2022; 11:e72904. https://doi.org/10.7554/eLife.72904, doi: 10.7554/eLife.72904, publisher: eLife Sciences Publications, Ltd.

Rutherford S, Kia SM, Wolfers T, Fraza C, Zabihi M, Dinga R, Berthet P, Worker A, Verdi S, Ruhe HG, Beckmann CF, Marquand AF. The normative modeling framework for computational psychiatry. Nature Protocols. 2022; p. 1–24. https://www.nature.com/articles/s41596-022-00696-5, doi: 10.1038/s41596-022-00696-5, publisher: Nature Publishing Group.

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. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. PloS One. 2017; 12(4):e0175683. doi: 10.1371/journal.pone.0175683.

Sanislow CA. RDoC at 10: changing the discourse for psychopathology. World Psychiatry. 2020; 19(3):311–312. https://onlinelibrary.wiley.com/doi/abs/10.1002/wps.20800, doi: 10.1002/wps.20800, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wps.20800.

738 **Schaefer A**, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, Eickhoff SB, Yeo BTT. Local-Global Par-
739 cellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex (New
740 York, NY). 2018; 28(9):3095–3114. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6095216/, doi: 10.1093/cer-
741 cor/bhx179.

742 **Shen X**, Tokoglu F, Papademetris X, Constable RT. Groupwise whole-brain parcellation from resting-state fMRI
743 data for network node identification. NeuroImage. 2013; 82:403–415. http://linkinghub.elsevier.com/retrieve/
744 pii/S1053811913005818, doi: 10.1016/j.neuroimage.2013.05.081.

745 **Shi D**, Li Y, Zhang H, Yao X, Wang S, Wang G, Ren K. Machine Learning of Schizophrenia Detection with Structural
746 and Functional Neuroimaging. Disease Markers. 2021; 2021:9963824. doi: 10.1155/2021/9963824.

747 **Siegel JS**, Mitra A, Laumann TO, Seitzman BA, Raichle M, Corbetta M, Snyder AZ. Data Quality Influences
748 Observed Links Between Functional Connectivity and Behavior. Cerebral Cortex (New York, NY). 2017;
749 27(9):4492–4502. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6410500/, doi: 10.1093/cercor/bhw253.

750 **Smith SM**, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, Filippini N, Watkins KE, Toro R, Laird AR, Beckmann
751 CF. Correspondence of the brain's functional architecture during activation and rest. Proceedings of the
752 National Academy of Sciences of the United States of America. 2009; 106(31):13040–13045. https://www.
753 ncbi.nlm.nih.gov/pmc/articles/PMC2722273/, doi: 10.1073/pnas.0905267106.

754 **Snelson E**, Rasmussen CE, Ghahramani Z. Warped Gaussian processes. In: *Proceedings of the 16th International*
755 *Conference on Neural Information Processing Systems* NIPS'03, MIT Press;. p. 337–344.

756 **Sprooten E**, Rasgon A, Goodman M, Carlin A, Leibu E, Lee WH, Frangou S. Addressing reverse inference in
757 psychiatric neuroimaging: Meta-analyses of task-related brain activation in common mental disorders. Hu-
758 man Brain Mapping. 2017; 38(4):1846–1864. https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23486, doi:
759 10.1002/hbm.23486, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.23486.

760 **Sripada C**, Angstadt M, Rutherford S, Kessler D, Kim Y, Yee M, Levina E. Basic Units of Inter-Individual Vari-
761 ation in Resting State Connectomes. Scientific Reports. 2019; 9(1):1900. https://www.nature.com/articles/
762 s41598-018-38406-5, doi: 10.1038/s41598-018-38406-5.

763 **Sripada C**, Angstadt M, Rutherford S, Taxali A, Shedden K. Toward a "treadmill test" for cognition: Im-
764 proved prediction of general cognitive ability from the task activated brain. Human Brain Mapping. 2020;
765 41(12):3186–3197. https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25007, doi: 10.1002/hbm.25007,
766 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.25007.

767 **Sripada C**, Rutherford S, Angstadt M, Thompson WK, Luciana M, Weigard A, Hyde LH, Heitzeg M. Prediction
768 of neurocognition in youth from resting state fMRI. Molecular Psychiatry. 2019; https://doi.org/10.1038/
769 s41380-019-0481-6, doi: 10.1038/s41380-019-0481-6.

770 **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,
771 Potkin SG, Preda A, Mathalon DH, Ford JM, Voyvodic J, Mueller BA, Belger A, et al. Multimodal neuromarkers
772 in schizophrenia via cognition-guided MRI fusion. Nature Communications. 2018; 9:3028. https://www.ncbi.
773 nlm.nih.gov/pmc/articles/PMC6072778/, doi: 10.1038/s41467-018-05432-w.

774 **Taxali A**, Angstadt M, Rutherford S, Sripada C. Boost in Test–Retest Reliability in Resting State fMRI with
775 Predictive Modeling. Cerebral Cortex. 2021; 31(6):2822–2833. https://doi.org/10.1093/cercor/bhaa390, doi:
776 10.1093/cercor/bhaa390.

777 **Tso IF**, Angstadt M, Rutherford S, Peltier S, Diwadkar VA, Taylor SF. Dynamic causal modeling of eye gaze
778 processing in schizophrenia. Schizophrenia Research. 2021; 229:112–121. https://www.sciencedirect.com/
779 science/article/pii/S092099642030551X, doi: 10.1016/j.schres.2020.11.012.

780 **Van Essen DC**, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, WU-Minn HCP Consortium.
781 The WU-Minn Human Connectome Project: an overview. NeuroImage. 2013; 80:62–79. doi:
782 10.1016/j.neuroimage.2013.05.041.

783 **Venkataraman A**, Whitford TJ, Westin CF, Golland P, Kubicki M. Whole Brain Resting State Functional Connec-
784 tivity Abnormalities in Schizophrenia. Schizophrenia Research. 2012; 139(1):7–12. https://www.ncbi.nlm.nih.
785 gov/pmc/articles/PMC3393792/, doi: 10.1016/j.schres.2012.04.021.

786 **Verdi S**, Marquand AF, Schott JM, Cole JH. Beyond the average patient: how neuroimaging models can address
787 heterogeneity in dementia. Brain. 2021; https://doi.org/10.1093/brain/awab165, doi: 10.1093/brain/awab165.

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. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020; 17(3):261–272. http://dx.doi.org/10.1038/s41592-019-0686-2, doi: 10.1038/s41592-019-0686-2.

Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E. An fMRI-Based Neurologic Signature of Physical Pain. New England Journal of Medicine. 2013; 368(15):1388–1397. https://doi.org/10.1056/NEJMoa1204471, doi: 10.1056/NEJMoa1204471, publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1204471.

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. Evidence for Network-Based Cortical Thickness Reductions in Schizophrenia. The American Journal of Psychiatry. 2019; 176(7):552–563. doi: 10.1176/appi.ajp.2019.18040380.

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. Quantifying Deviations of Brain Structure and Function in Major Depressive Disorder Across Neuroimaging Modalities. JAMA Psychiatry. 2022; 79(9):879–888. https://doi.org/10.1001/jamapsychiatry.2022.1780, doi: 10.1001/jamapsychiatry.2022.1780.

Wolfers T, Arenas AL, Onnink AMH, Dammers J, Hoogman M, Zwiers MP, Buitelaar JK, Franke B, Marquand AF, Beckmann CF. Refinement by integration: aggregated effects of multimodal imaging markers on adult ADHD. Journal of Psychiatry & Neuroscience : JPN. 2017; 42(6):386–394. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5662460/, doi: 10.1503/jpn.160240.

Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neuroscience & Biobehavioral Reviews. 2015; 57:328–349. http://www.sciencedirect.com/science/article/pii/S0149763415002018, doi: 10.1016/j.neubiorev.2015.08.001.

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. Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. JAMA Psychiatry. 2018; 75(11):1146–1155. https://doi.org/10.1001/jamapsychiatry.2018.2467, doi: 10.1001/jamapsychiatry.2018.2467.

Wolfers T, Rokicki J, Alnæs D, Berthet P, Agartz I, Kia SM, Kaufmann T, Zabihi M, Moberget T, Melle I, Beckmann CF, Andreassen OA, Marquand AF, Westlye LT. Replicating extensive brain structural heterogeneity in individuals with schizophrenia and bipolar disorder. Human Brain Mapping. 2021; 42(8):2546–2555. https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25386, doi: https://doi.org/10.1002/hbm.25386, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.25386.

Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. Nature Neuroscience. 2017; 20(3):365–377. http://www.nature.com/articles/nn.4478, doi: 10.1038/nn.4478.

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. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of Neurophysiology. 2011; 106(3):1125–1165. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3174820/, doi: 10.1152/jn.00338.2011.

Yu Q, Allen EA, Sui J, Arbabshirani MR, Pearlson G, Calhoun VD. Brain connectivity networks in schizophrenia underlying resting state functional magnetic resonance imaging. Current topics in medicinal chemistry. 2012; 12(21):2415–2425. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4429862/.

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. Fractionating autism based on neuroanatomical normative modeling. Translational Psychiatry. 2020; 10(1):1–10. https://www.nature.com/articles/s41398-020-01057-0, doi: 10.1038/s41398-020-01057-0, number: 1 Publisher: Nature Publishing Group.

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. Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. 2019; 4(6):567–578. https://www.sciencedirect.com/science/article/pii/S245190221830329X, doi: 10.1016/j.bpsc.2018.11.013.

842 **Zhang J**, Kucyi A, Raya J, Nielsen AN, Nomi JS, Damoiseaux JS, Greene DJ, Horovitz SG, Uddin LQ, Whitfield-
843 Gabrieli S. What have we really learned from functional connectivity in clinical populations? Neu-
844 roImage. 2021; 242:118466. https://www.sciencedirect.com/science/article/pii/S1053811921007394, doi:
845 10.1016/j.neuroimage.2021.118466.

Functional MRI Acquisition Parameters

In the HCP study, four runs of resting state fMRI data (14.5 minutes 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, 2mm isotropic voxels, FOV=208×180mm, 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, 0.7mm isotropic voxels, FOV=224mm, 256 sagittal slices) and T2 weighted scans were acquired with a SPACE sequence (TR=3200ms, TE=565ms, 0.7mm isotropic voxels, FOV=224mm, 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 = 2000 ms, TE = 29 ms, FOV = 64x64, 32 slices in axial plane interleaved multi slice series ascending, voxel size = 3x3x4 mm3). The University of Michigan SchizGaze study was collected in two phases with different parameters but using the same MRI machine (3.0 T GE MR 750 Discovery scanner). In SchizGaze1 (N=47), functional images were acquired with a T2*-weighted, reverse spiral acquisition sequence (TR = 2000 ms, 240 volumes (8 minutes), 3mm 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 = 800 ms, 453 volumes (6 minutes), 2.4mm 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 3, including the sample size, age (mean and standard deviation), and sex distribution of each site. Many sites were pulled from publicly available data sets including ABCD, CAMCAN, CMI-HBN, HCP-Aging, HCP-Development, HCP-Early Psychosis, HCP-Young Adult, NKI-RS, OpenNeuro, PNC, and UKBiobank. For data sets 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 data sets (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 to create a multi-site transfer data set.

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 between network connectivity pair (Yeo-17 and Smith-10) from a vector of covariates (age, sex, site, meanFD). For a detailed mathematical description see (***Fraza et al., 2021***). Briefly, for each brain region of interest, y is predicted as:

$$y = \omega^{\intercal}\varphi(x) + \epsilon \tag{4}$$

Where $\omega^{\intercal}$ is the estimated weight vector, $\varphi(x)$ 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 $\epsilon = \eta(\theta, \beta)$ a Gaussian noise distribution with mean zero and noise precision term β (the inverse variance). A likelihood warping approach (***Rios and Tobar, 2019***; ***Snelson et al., ????***) was used to model non-Gaussian effects. This involves 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 employed a 'sinarcsinsh' warping function, which is equivalent to the SHASH distribution commonly used in the generalized additive modeling literature (***Jones and Pewsey, 2009***) and which we have found to perform well in prior work (***Dinga et al., 2021***; ***Fraza et al., 2021***). Site variation was modeled using fixed effects, which we have shown in prior work provides relatively good performance (***Kia et al., 2021***), although random effects for site may provide additional flexibility at higher computational cost. A fast numerical optimization algorithm was used to optimize hyperparameters ('Powell'). Computational complexity of hyperparameter optimization was controlled by minimizing the negative log likelihood. Deviation scores (Z-scores) are calculated for the n_{th} subject, and d_{th} brain area, in the test set as:

$$Z_{n_d} = \frac{y_{n_d} - \hat{y_{n_d}}}{\sqrt{(\theta_d)^2 + (\theta_{*_d})^2}} \tag{5}$$

Where y_{n_d} is the true response, $\hat{y_{n_d}}$ is the predicted mean, θ_d^2 is the estimated noise variance (reflecting uncertainty in the data), and $\theta_{*_d}^2$ 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 (equation 5) which measures how well the shape of the regression function matches the data (***Dinga et al., 2021***).

Appendix 0 Table 3. Functional normative model train/test demographics per site.

	Train			Test		
Site	N	Sex (F/M)%	Age (m, s.d)	N	Sex (F/M)%	Age (m, s.d)
ABCD_01	60	48.33, 51.67	9.87, 0.58	73	56.16, 43.84	9.95/0.61
ABCD_02	244	47.54, 52.46	10.12, 0.64	258	47.29, 52.71	10.1, 0.62
ABCD_03	282	47.87, 52.13	9.87, 0.62	260	50, 50	9.91, 0.61
ABCD_04	258	49.61, 50.39	9.91, 0.64	268	50.37, 49.63	9.77, 0.64
ABCD_05	161	62.11, 37.89	9.85, 0.63	144	42.36, 57.64	9.96, 0.63
ABCD_06	228	54.82, 45.18	9.98, 0.58	240	48.75, 51.25	10.02, 0.59
ABCD_07	128	50.78, 49.22	9.86, 0.64	128	42.97, 57.03	9.93, 0.61
ABCD_08	113	50.44, 49.56	9.98, 0.63	104	45.19, 54.81	10.1, 0.59
ABCD_09	173	48.55, 51.45	10.05, 0.59	175	56, 44	9.89, 0.60
ABCD_10	187	48.66, 51.34	9.88, 0.62	223	45.29, 54.71	9.93, 0.64
ABCD_11	192	51.04, 48.96	9.88, 0.65	173	51.45, 48.55	9.79, 0.62
ABCD_12	70	50, 50	9.85, 0.60	68	1.47, 54.41	9.95, 0.55
ABCD_13	209	51.67, 48.33	9.84, 0.61	191	52.88, 47.12	9.84, 0.60
ABCD_14	286	48.6, 51.4	10.24, 0.51	220	45.45, 54.55	10.2, 0.54
ABCD_15	138	55.07, 44.93	9.94, 0.62	149	45.64, 54.36	10.0, 0.58
ABCD_16	458	46.29, 53.71	9.89, 0.64	462	42.64, 48.7	9.90, 0.66
ABCD_17	204	54.9, 45.1	9.84, 0.61	221	41.63, 58.37	9.87, 0.65

ABCD_18	97	37.11, 62.89	9.91, 0.67	109	53.21, 46.79	10.0, 0.60
ABCD_19	187	55.08, 44.92	10.12, 0.55	205	52.68, 47.32	10.1, 0.54
ABCD_20	278	50.72, 49.28	10.05, 0.48	259	54.44, 45.56	10.1, 0.50
ABCD_21	212	44.81, 55.19	9.97, 0.63	238	49.58, 50.42	9.94, 0.61
AOMIC_PIPO1	162	58.64, 41.36	22.2, 1.8	41	53.66, 46.34	22.4, 1.7
AOMIC_PIPO2	166	59.64, 40.36	22.2, 1.7	41	46.34, 53.66	22.1, 2.2
CAMCAN	495	49.49, 50.51	53.2, 18.3	124	54.84, 45.16	55.3, 20.0
CMI-HBN_CBIC	133	37.59, 62.41	11.9, 3.4	33	48.48, 51.52	11.7, 3.6
CMI-HBN_RU	74	37.84, 62.16	11.6, 3.6	18	38.89, 61.11	10.9, 3.5
CNP-35343.0	79	44.3, 55.7	31.1, 9.1	19	52.63, 47.37	32.1, 7.4
CNP-35426.0	18	44.44, 55.56	31.1, 8.2	4	75, 25	34.8, 10.9
HCP_A_MGH	130	51.54, 48.46	62.1, 16.1	33	51.52, 48.48	59.1, 15.6
HCP_A_UCLA	118	56.78, 43.22	55.9, 13.1	30	60, 40	59.3, 15.8
HCP_A_UM	164	56.71, 43.29	61.9, 17.0	41	48.78, 51.22	60.8, 16.2
HCP_A_WU	167	61.08, 38.92	61.0, 15.4	42	52.38, 47.62	60.2, 15.2
HCP_D_MGH	137	54.01, 45.99	14.7, 3.8	34	55.88, 44.12	13.1, 3.4
HCP_D_UCLA	82	50, 50	14.6, 3.7	21	23.81, 28.57	14.7, 4.0
HCP_D_UM	99	54.55, 44.44	13.8, 3.7	24	66.67, 37.5	13.7, 3.9
HCP_D_WU	94	48.94, 51.06	14.5, 3.9	23	52.17, 47.83	15.2, 4.2
HCP_YA	500	51.8, 48.2	28.5, 3.8	501	54.49, 45.51	28.9, 3.6
NKI-RS	136	48.53, 51.47	21.1, 6.5	34	50, 50	17.8, 6.5
PNC	630	56.03, 43.97	14.6, 3.3	158	43.04, 56.96	14.6, 8.0
ukb	6924	55.37, 44.63	62.4, 7.5	1732	55.25, 44.75	63.3, 7.5
UMich_IMPs	235	52.77, 47.23	12.9, 3.4	59	54.24, 45.76	12.5, 3.6

Appendix 0 Table 4. Surface area normative model demographics per site

Site	N	Age (m, s.d.)	Sex (F/M) %
ABCD_01	388	9.90, 0.62	51.29, 48.71
ABCD_02	542	10.1, 0.62	46.49, 53.51
ABCD_03	569	9.88, 0.66	47.1, 52.9
ABCD_04	631	9.82, 0.71	48.65, 51.35
ABCD_05	345	9.89, 0.63	51.59, 48.41
ABCD_06	564	9.94, 0.59	50.71, 49.29
ABCD_07	325	9.87, 0.62	47.08, 52.92
ABCD_08	336	9.95, 0.62	47.62, 52.38
ABCD_09	407	9.96, 0.61	49.14, 50.86
ABCD_10	575	9.86, 0.62	48.7, 51.3
ABCD_11	414	9.82, 0.62	49.76, 50.24
ABCD_12	161	9.88, 0.59	47.83, 52.17
ABCD_13	555	9.82, 0.59	49.91, 50.09
ABCD_14	583	10.2, 0.57	45.8, 54.2
ABCD_15	396	9.90, 0.60	44.95, 55.05
ABCD_16	921	9.90, 0.65	44.95, 55.05
ABCD_17	557	9.82, 0.63	47.94, 52.06
ABCD_18	341	9.91, 0.63	46.92, 53.08
ABCD_19	534	10.1, 0.55	50.94, 49.06

ABCD_20	640	10.0, 0.49	50, 50
ABCD_21	491	9.91, 0.62	45.21, 54.79
ABCD_22	35	10.2, 0.55	62.86, 37.14
ATT	31	23.0, 1.88	9.68, 90.32
ATV	77	22.7, 1.98	77.92, 22.08
CAMCAN	647	54.2, 18.6	50.85, 49.15
CIN	136	52.4, 15.5	36.76, 63.24
CMI-RU	563	10.3, 3.49	34.81, 65.19
CMI-SI	341	11.2, 3.83	43.11, 56.89
CNP-35343.0	153	33.3, 9.38	47.71, 52.29
CNP-35426.0	79	33.3, 9.24	32.91, 67.09
COI	193	49.4, 13.5	61.14, 38.86
ds001734	108	25.5, 3.59	55.56, 44.44
ds002236	86	11.5, 2.04	44.19, 55.81
ds002330	65	26.2, 4.30	55.38, 44.62
ds002345	207	21.7, 4.71	63.29, 36.71
ds002731	59	21.3, 1.45	47.46, 52.54
ds002837	86	26.7, 10.1	48.84, 51.16
HCP-Aging_MGH	171	59.8, 15.5	50.29, 49.71
HCP-Aging_UCLA	124	53.3, 12.8	57.26, 42.74
HCP-Aging_UMinn	204	61.6, 16.3	58.82, 41.18
HCP-Aging_WashU	178	58.5, 13.8	62.92, 37.08
HCP-Dev_MGH	216	13.8, 3.87	50.46, 49.54
HCP-Dev_UCLA	127	14.1, 3.82	48.82, 51.18
HCP-Dev_UMinn	156	13.3, 3.64	54.49, 45.51
HCP-Dev_WashU	154	14.0, 3.87	48.7, 51.3
HCP-EP_BWH	31	22.6, 4.00	32.26, 67.74
HCP-EP_IU	84	23.2, 3.82	39.29, 60.71
HCP-EP_McL	44	24.1, 3.56	43.18, 56.82
HCP-EP_MGH	21	24.1, 5.44	28.57, 71.43
HCP-YA	1113	28.8, 3.70	54.45, 45.55
HKH	62	45.1, 10.5	48.39, 51.61
HRC	65	41.4, 11.5	70.77, 29.23
HUH	124	38.7, 13.3	50.81, 49.19
IXI	581	49.5, 16.7	56.28, 43.72
KTT	121	32.4, 10.3	38.84, 61.16
KUT	220	38.0, 13.1	43.64, 56.36
NKI	482	42.6, 21.2	63.9, 36.1
NKN	9	63.6, 18.5	44.44, 55.56
Oasis3	2044	70.4, 9.51	42.37, 57.63
PNC	1378	14.2, 3.51	50.87, 49.13
SWA	234	31.4, 8.75	14.53, 85.47
SWU_SLIM	569	20.1, 1.27	56.24, 43.76
TOP	823	33.2, 10.2	47.14, 52.86
ukb-11025.0	16132	62.5, 7.50	51.83, 48.17
ukb-11026.0	658	65.3, 7.37	54.86, 45.14
ukb-11027.0	3880	63.7, 7.46	53.92, 46.08
UTO	351	35.4, 14.6	45.58, 54.42